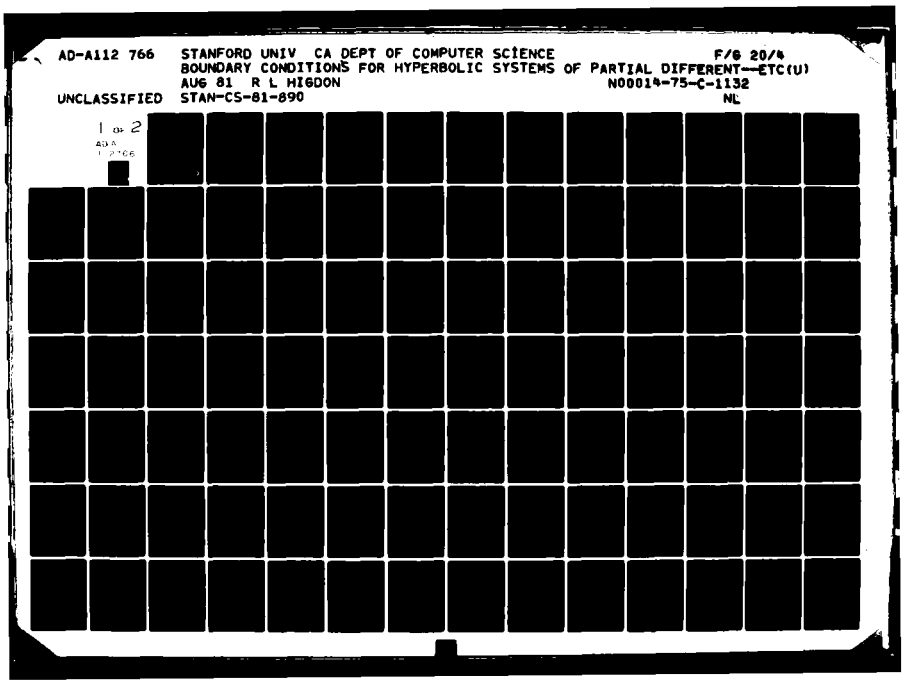
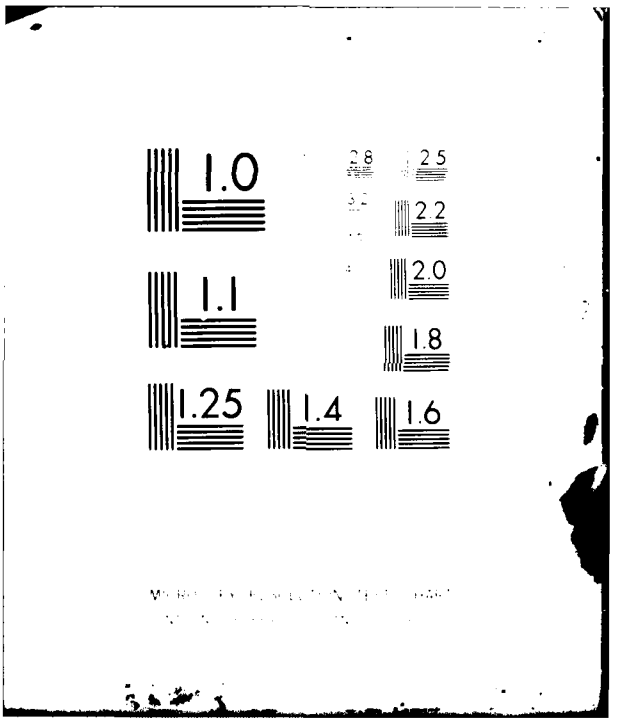

AD-A112 766
STANFORD UNIV CA DEPT OF COMPUTER SCIENCE
F/G 20/4
BOUNDARY CONDITIONS FOR HYPERBOLIC SYSTEMS OF PARTIAL DIFFERENT--ETC(U)
AUG 81 R L HIGDON
N00014-75-C-1132
UNCLASSIFIED
STAN-CS-81-890
NL
1 of 2
AD A
112766


1.0
2.8
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

August 1981

Report. No. STAN-CS-81-890

Boundary Conditions for Hyperbolic Systems of Partial Differential Equations Having Multiple Time Scales

by

Robert L. Higdon

Department of Computer Science

Stanford University
Stanford, CA 94305

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER STAN-CS-81-890	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* Boundary Conditions for Hyperbolic Systems of Partial Differential Equations Having Multiple Time Scales		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Robert L. Higdon		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-1132
9. PERFORMING ORGANIZATION NAME AND ADDRESS Stanford University Department of Computer Science Stanford, California 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61153N RR 014-02-01 NR 044-492
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Mathematics Program Arlington, VA 22217		12. REPORT DATE August 1981 13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Hyperbolic Partial Differential Equations
Boundary Conditions
Multiple Time Scales
Computational Methods

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

This paper is concerned with linear hyperbolic systems of partial differential equations for which certain of the associated propagation speeds are a great deal larger than the other propagation speeds. In certain cases the fast modes allowed by such a system are not present in the true physical solution. Yet the fact that such modes are allowed means that when one tries to compute a numerical solution to an initial-boundary value problem, the errors generated can propagate quite rapidly. In particular, when the

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

boundary data used for the computation are less accurate than the initial data, the fast modes can cause a rapid contamination of the calculation in the interior. To prevent this, one would like to have boundary conditions which prevent fast waves from entering the region. The goal of this paper is to find such conditions.

The situation described here is often encountered when equations of gas dynamics are used to model the behavior of the earth's atmosphere. This is the physical problem which motivated this study.

In order to find the desired boundary conditions, we first transform the given system to an approximate diagonal form in such a way that each of the new dependent variables can be identified as a slow, incoming fast, or outgoing fast component of the solution. We then find local boundary conditions which suppress the incoming fast part. Pseudo-differential operators are used throughout the entire process. The effects of these boundary conditions are analyzed using methods from the theory of propagation of singularities for linear partial differential equations.

This process has been worked out in detail for a model problem in one space dimension and for the linearized shallow water equations, a system in two space dimensions. We have included the results of some numerical calculations which demonstrate the effectiveness of the boundary conditions.

INSTRUCTIONS FOR PREPARATION OF REPORT DOCUMENTATION PAGE

RESPONSIBILITY. The controlling DoD office will be responsible for completion of the Report Documentation Page, DD Form 1473, in all technical reports prepared by or for DoD organizations.

CLASSIFICATION. Since this Report Documentation Page, DD Form 1473, is used in preparing announcements, bibliographies, and data banks, it should be unclassified if possible. If a classification is required, identify the classified items on the page by the appropriate symbol.

COMPLETION GUIDE

General. Make Blocks 1, 4, 5, 6, 7, 11, 13, 15, and 16 agree with the corresponding information on the report cover. Leave Blocks 2 and 3 blank.

Block 1. Report Number. Enter the unique alphanumeric report number shown on the cover.

Block 2. Government Accession No. Leave Blank. This space is for use by the Defense Documentation Center.

Block 3. Recipient's Catalog Number. Leave blank. This space is for the use of the report recipient to assist in future retrieval of the document.

Block 4. Title and Subtitle. Enter the title in all capital letters exactly as it appears on the publication. Titles should be unclassified whenever possible. Write out the English equivalent for Greek letters and mathematical symbols in the title *(see "Abstracting Scientific and Technical Reports of Defense-sponsored RDT/E," AD-667 000)*. If the report has a subtitle, this subtitle should follow the main title, be separated by a comma or semicolon if appropriate, and be initially capitalized. If a publication has a title in a foreign language, translate the title into English and follow the English translation with the title in the original language. Make every effort to simplify the title before publication.

Block 5. Type of Report and Period Covered. Indicate here whether report is interim, final, etc., and, if applicable, inclusive dates of period covered, such as the life of a contract covered in a final contractor report.

Block 6. Performing Organization Report Number. Only numbers other than the official report number shown in Block 1, such as series numbers for in-house reports or a contractor/grantee number assigned by him, will be placed in this space. If no such numbers are used, leave this space blank.

Block 7. Author(s). Include corresponding information from the report cover. Give the name(s) of the author(s) in conventional order *(for example, John R. Doe or, if author prefers, J. Robert Doe)*. In addition, list the affiliation of an author if it differs from that of the performing organization.

Block 8. Contract or Grant Number(s). For a contractor or grantee report, enter the complete contract or grant number(s) under which the work reported was accomplished. Leave blank in in-house reports.

Block 9. Performing Organization Name and Address. For in-house reports enter the name and address, including office symbol, of the performing activity. For contractor or grantee reports enter the name and address of the contractor or grantee who prepared the report and identify the appropriate corporate division, school, laboratory, etc., of the author. List city, state, and ZIP Code.

Block 10. Program Element, Project, Task Area, and Work Unit Numbers. Enter here the number code from the applicable Department of Defense form, such as the DD Form 1498, "Research and Technology Work Unit Summary" or the DD Form 1634, "Research and Development Planning Summary," which identifies the program element, project, task area, and work unit or equivalent under which the work was authorized.

Block 11. Controlling Office Name and Address. Enter the full, official name and address, including office symbol, of the controlling office. *(Equates to funding/sponsoring agency. For definition see DoD Directive 5200.20, "Distribution Statements on Technical Documents.")*

Block 12. Report Date. Enter here the day, month, and year or month and year as shown on the cover.

Block 13. Number of Pages. Enter the total number of pages.

Block 14. Monitoring Agency Name and Address *(if different from Controlling Office)*. For use when the controlling or funding office does not directly administer a project, contract, or grant, but delegates the administrative responsibility to another organization.

Blocks 15 & 15a. Security Classification of the Report: Declassification/Downgrading Schedule of the Report. Enter in 15 the highest classification of the report. If appropriate, enter in 15a the declassification/downgrading schedule of the report, using the abbreviations for declassification/downgrading schedules listed in paragraph 4-207 of DoD 5200.1-R.

Block 16. Distribution Statement of the Report. Insert here the applicable distribution statement of the report from DoD Directive 5200.20, "Distribution Statements on Technical Documents."

Block 17. Distribution Statement *(of the abstract entered in Block 20, if different from the distribution statement of the report)*. Insert here the applicable distribution statement of the abstract from DoD Directive 5200.20, "Distribution Statements on Technical Documents."

Block 18. Supplementary Notes. Enter information not included elsewhere but useful, such as: Prepared in cooperation with . . . Translation of *(or by)* . . . Presented at conference of . . . To be published in . . .

Block 19. Key Words. Select terms or short phrases that identify the principal subjects covered in the report, and are sufficiently specific and precise to be used as index entries for cataloging, conforming to standard terminology. The DoD "Thesaurus of Engineering and Scientific Terms" *(TEST)*, AD-672 000, can be helpful.

Block 20. Abstract. The abstract should be a brief *(not to exceed 200 words)* factual summary of the most significant information contained in the report. If possible, the abstract of a classified report should be unclassified and the abstract to an unclassified report should consist of publicly-releasable information. If the report contains a significant bibliography or literature survey, mention it here. For information on preparing abstracts see "Abstracting Scientific and Technical Reports of Defense-Sponsored RDT&E," AD-667 000.

NR 044-402

BOUNDARY CONDITIONS FOR HYPERBOLIC SYSTEMS OF PARTIAL DIFFERENTIAL EQUATIONS HAVING MULTIPLE TIME SCALES

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Robert Lynn Higdon

August 1981

ABSTRACT

This paper is concerned with linear hyperbolic systems of partial differential equations for which certain of the associated propagation speeds are a great deal larger than the other propagation speeds. In certain cases the fast modes allowed by such a system are not present in the true physical solution. Yet the fact that such modes are allowed means that when one tries to compute a numerical solution to an initial-boundary value problem, the errors generated can propagate quite rapidly. In particular, when the boundary data used for the computation are less accurate than the initial data, the fast modes can cause a rapid contamination of the calculation in the interior. To prevent this, one would like to have boundary conditions which prevent fast waves from entering the region. The goal of this paper is to find such conditions.

The situation described here is often encountered when equations of gas dynamics are used to model the behavior of the earth's atmosphere. This is the physical problem which motivates this study.

In order to find the desired boundary conditions, we first transform the given system to an approximate diagonal form in such a way that each of the new dependent variables can be identified as a slow, incoming fast, or outgoing fast component of the solution. We then find local boundary conditions which suppress the incoming fast part. Pseudo-differential operators are used throughout the entire process. The effects of these boundary conditions are analyzed using methods from the theory of propagation of singularities for linear partial differential equations.

This process has been worked out in detail for a model problem in one space dimension and for the linearized shallow water equations, a system in two space dimensions. We have included the results of some numerical calculations which demonstrate the effectiveness of the boundary conditions.

ACKNOWLEDGMENTS

I would like to thank my advisor, Joseph Oliger, for the advice and encouragement he has given me and for making it very easy for me to pursue a degree program involving both the Department of Mathematics and the Department of Computer Science. I also thank his students in the Department of Computer Science for providing technical assistance in carrying out the numerical computations which are presented in this thesis. For part of the time when I worked on this thesis I was supported through Professor Oliger's contract N00014-75-C-1132 with the Office of Naval Research. Numerical computations were performed through the computing services of the Stanford Linear Accelerator Center.

My stay at Stanford has been very pleasant and stimulating. This is due partly to the people already mentioned. I should also thank the graduate students, faculty, and staff of the Department of Mathematics.

I extend special thanks to my parents for the support and encouragement which they have given me over the years.

The manuscript was typed by Beth Arrington. I thank her for the speed and accuracy with which she performed this task.

TABLE OF CONTENTS

Chapter 1. Introduction 1

Chapter 2. The Problem in One Space Dimension 10

2.1. General Remarks 10

2.2. Outline of a Method for Uncoupling Systems of Equations 14

2.3. Transformations in Time 20

2.4. A More Complete Treatment of the Uncoupling Method 25

2.5. Boundary Conditions 31

2.6. Estimates of the Size of the Fast Part of the Solution: Outline and Physical Interpretation . . . 40

2.7. Estimates of the Size of the Fast Part of the Solution: A More Complete Treatment 48

2.8. Numerical Computations 58

Chapter 3. The Problem in Several Space Dimensions 61

3.1. Properties of the Principal Symbol 61

3.2. Outline of the Uncoupling Process 72

3.3. A Perturbation Lemma 81

Chapter 4. An Example in Two Space Dimensions 84

4.1. Uncoupling the System 84

4.2. Boundary Conditions 99

4.3. Effects of Orthogonal Changes of Spatial Coordinates 108

4.4. Numerical Computations 117

Appendix. Properties of Pseudo Differential Operators 126

Bibliography 131

CHAPTER 1

INTRODUCTION

Hyperbolic partial differential equations are characterized by the fact that in a certain sense they propagate information at finite speed. For first order hyperbolic systems there may be several such propagation speeds, each corresponding to an eigenvalue of the principal symbol of the system. In this paper we will consider systems for which the various speeds can have substantially different magnitudes. Such systems are sometimes said to have "multiple time scales."

Examples of these systems arise in the study of fluid dynamics. For such systems there are certain propagation modes related to the movement of the fluid, and there are certain other modes which have a different physical interpretation. For the Euler equations of gas dynamics these other modes are associated with the movement of sound waves, and for the shallow water equations they are related to the movement of gravity waves. If these waves move at speeds which are considerably greater than the rate of flow of the fluid, then these systems have two time scales.

The work presented here is concerned with a certain difficulty which can arise when one tries to compute numerical approximations to the solutions of such systems. The physical problem which motivates this study is the use of hyperbolic systems to model the behavior of the earth's atmosphere.

The specific situation is illustrated in Figure 1.1. This figure shows the domain of definition for an initial-boundary value problem for a hyperbolic system in one space dimension. In this case the spatial region is an interval I, and the system is to be studied for time $t > 0$. The restriction to one space dimension is made solely for the purpose of keeping the picture simple. In order to define a well-posed problem on this space-time domain, it is necessary to specify values for the solution at time $t = 0$, and it is also necessary to specify certain conditions at the boundary of I throughout all positive time. In specific situations the necessary data are taken from physical measurements.

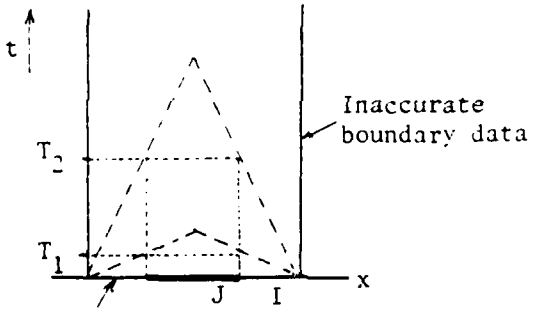


Figure 1.1

In certain meteorological problems the boundary data available for a numerical computation are often considered to be substantially less accurate than the available initial data. The reasons for this will be discussed a little later. This state of affairs is unfortunate, since the inaccuracies in the boundary data will generate comparable

inaccuracies in the interior, thereby wasting the extra accuracy contained in the initial data. Our goal is to control this contamination as much as possible.

This sort of control is feasible because hyperbolic systems can propagate information only at finite speed. For any subregion J of the given region I, there is a certain period of time after $t = 0$ during which the solution in J cannot be influenced by the boundary data. However, it is inevitable that the boundary values will eventually influence the solution in J and thereby reduce the accuracy of the computed solution in that region. The question is how long this takes. We have noted that the systems in question can allow both slow and fast propagation speeds. These are illustrated by the characteristic lines appearing in Figure 1.1. If the boundary data influence the interior at the fast speed, then in the region J the solution is accurate up to the time T_1 indicated in the figure. If the boundary data move in at the slow speed, then the computed solution is accurate up to time T_2. In the meteorological problem these times can easily differ by a factor of five to ten. It would therefore be worthwhile to prevent this contamination from taking place at the faster speed.

In order to do this we will try to find boundary conditions which prevent rapidly moving waves from entering the given spatial region. We will try to identify, in some sense, the portion of the solution which is entering the region at the fast speed, and we will then attempt to set this part of the solution equal to zero at the boundary.

For the meteorological problem this would accomplish what we want. The crucial feature of this problem is that boundary conditions of this type are entirely realistic. In this problem the fast modes in the system correspond to the motion of sound waves or gravity waves. The amount of energy contained in such waves is insignificant compared to the other forms of energy in the atmosphere, so for practical purposes the fast part of the exact solution is in fact equal to zero. The only reason that the fast modes can cause any trouble is that a numerical computation can introduce errors which have nothing to do with the exact solution. These errors can therefore be propagated by all of the modes in the system. Because the fast part can consist only of errors, it is entirely reasonable to try to suppress this part of the solution. We will not attempt to prevent the propagation of error at the slow speed, since it is not realistic to assume that the slow part of the solution is equal to zero. We will instead accept the fact that on any subregion the computed solution will eventually suffer reduced accuracy due to the effect of the boundary data.

This discussion has been based on the fact that there are certain modes allowed by the system which are not present in the true physical solution. These modes do not contribute to a description of the physical situation, but they do cause problems when we try to compute numerical approximations to the solution of the system. We have mentioned one such problem, and we will mention another a little later. It might seem that we could best deal with these problems by modifying the system of differential equations so as to prohibit solutions containing rapidly moving waves. This would certainly eliminate the

problems, and it would also be physically reasonable. In meteorology this process is known as "filtering". However, there are no known filtered systems which are mathematically well-behaved and which are sufficiently accurate models of the atmosphere to be useful in meteorological calculations.

There is a partially filtered system, known as the "primitive equations", which is currently being used for such calculations. This system is derived from the Euler equations of gas dynamics, and it is based on the assumption that the atmosphere is in hydrostatic balance. This assumption prevents vertically moving sound waves from appearing in the solution. Unfortunately, this sytem does not have certain desirable mathematical properties. The system is not hyperbolic, and it has been shown by Oliger and Sundström [7] that it is not possible to find local boundary conditions at open boundaries which lead to well-posed initial-boundary value problems for this system. In current practice a diffusion term is added to the system to make it parabolic, and values for all components are then prescribed at the boundary. This leads to a well-posed problem, but it also reduces the accuracy of the solution which is computed.

Because of the difficulties involved with finding a suitable filtered system, it may be desirable to use the unaltered Euler equations of gas dynamics for meteorological computations. The work presented in this paper is concerned with one of the difficulties which can arise when we try to do this.

There is another difficulty which can arise in this situation.

Because of the Courant-Friedrichs-Lewy condition, the fast modes in the system can impose a severe restriction on the permissible time step for stable explicit difference approximations to the differential equation. In general, this would present us with the choice of either using an implicit difference method or an explicit method with very short time steps. Both choices would be undesirable because of the computational expense which would be involved. However, some recent work by Kreiss [5] has made it possible to avoid this problem by choosing a suitable set of initial data. He has shown that certain problems of instability can be avoided if we smooth our given data so that certain elliptic equations in the spatial variables are satisfied at the initial time. This makes it possible to use an explicit difference scheme having a reasonable time step.

We still need to discuss the reason why the boundary data available for certain meteorological computations are considered to be substantially less accurate than the available initial data. This situation arises in limited area computations which are used to predict local atmospheric phenomena. Such computations are made necessary by the size of the earth's atmosphere. If we try to compute the solution to a system of equations over the entire atmosphere, then it will be necessary to use an extremely coarse grid for the difference equations. Otherwise, the computation would be too lengthy for present-day computing machines. In current practice the grid spacing for global atmospheric computations is roughly one interval per two and a half degrees

of latitude and longitude. Such computations can give useful information about global phenoman, but the grid spacing is too coarse for predicting local phenomena.

It is common practice to perform additional computations over smaller regions with finer meshes so that these local phenomena can be resolved. For such a computation the spatial region is a cyclinder in the atmosphere which is bounded by the earth's surface, the top of the atmosphere, and an artificial computational boundary. This artificial boundary merely defines the edge of the computation and represents nothing physical. The situation is illustrated schematically in Figure 1.2. The global computation is represented by the coarse grid, and the local computation is represented by the finer grid in the interval I. It is necessary to find suitable initial data and boundary

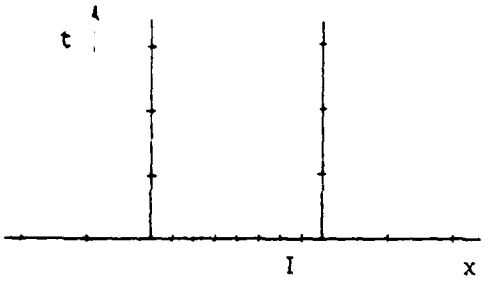


Figure 1.2

data for this computation. We can safely assume that we can find accurate initial data, since we would perform local computations only over a populated region where there is a dense network of observation stations which are capable of accurate measurements. The

problem is in finding suitable boundary data when we are trying to predict future weather patterns. For such data we would have to use the results of our global prediction. This computation is made on a mesh which is much coarser than that of the local computation, so the results of this computation cannot be considered as accurate as the initial data which are available. This is one source of the inaccuracies in the boundary data which we have been discussing.

We now outline the contents of this paper. We will consider initial-boundary value problems for hyperbolic systems having two time scales. Our goal is to find boundary conditions which suppress the part of the solution which would enter the given spatial domain at the fast speed. Although the systems of real interest are quasi-linear, we will consider only linearized systems. Our hope is that a study of such systems can eventually lead to useful boundary conditions for the nonlinear problem.

Our basic method is to diagonalize the system in such a way that each of the new dependent variables can be identified as a slow, incoming fast, or outgoing fast component of the solution. We will then attempt to set the incoming fast part of the solution equal to zero at the boundary. Our methods will rely heavily on the use of *pseudo* differential operators. In the Appendix we will define a common class of such operators, and we will state without proof some of their basic properties.

In Chapter 2 we will discuss the problem for hyperbolic systems in one space dimension, and in Chapter 3 we will generalize the methods of Chapter 2 to problems in several space dimensions. We will use these

techniques in Chapter 4 to derive boundary conditions for the linearized shallow water equations. The results of some numerical computations will be included.

The work presented here is related to some work by Engquist and Majda on absorbing boundary conditions. In [1] they suggested some methods for constructing such conditions both for scalar wave equations and for first order hyperbolic systems. Some of the methods which we will use here resemble, in rough outline, the ideas they proposed for hyperbolic systems. Their ideas for scalar wave equations are developed in detail in [2].

CHAPTER 2

THE PROBLEM IN ONE SPACE DIMENSION

In this chapter we consider the situation for hyperbolic systems in one space variable. The problems of real interest occur in more than one space dimension, but certain features of the general problem can be seen in this simpler special case.

2.1 General Remarks

We will consider the hyperbolic system

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a & \\ & b \end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} \tag{2.1}$$

for $0 < x < 1$, $t > 0$. This can also be written $w_t = Aw_x + Cw$, where $w = (u,v)^T \in \mathbf{R}^2$. The entries in A and C are functions of x and t.

In order to simplify the notation we have chosen a system having two scalar components. It can be seen easily that the ideas presented in this chapter work equally well for systems having several components.

There is no loss of generality in assuming that A is diagonal. The system is hyperbolic, so A has real eigenvalues and a complete set of eigenvectors. If A is not diagonal, then a suitable similarity transformation and change of dependent variables can be made to bring the system to diagonal form.

We assume $|a| << |b|$ and $a < 0$, $b < 0$. The first assumption guarantees the presence of propagation speeds having substantially

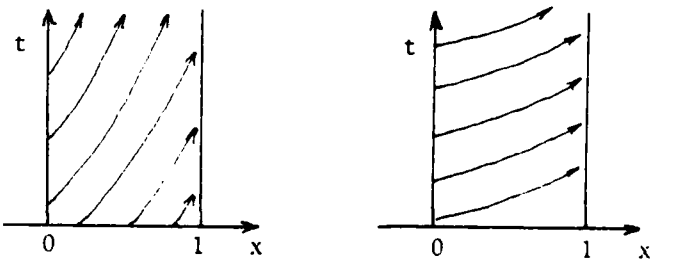
different magnitudes. The second assumption is made for the sake of definiteness. It also contains the assumption that $\det A \neq 0$, i.e., that the boundary is noncharacteristic.

The problem is to identify the "fast" part of the solution of the system and then find boundary conditions which suppress this as much as possible. To some degree this can be done by considering the usual method of characteristics for constructing the solution of the system. Suppose first that the matrix C is diagonal. The system (2.1) then uncouples into two independent equations

$$u_t = au_x + c_{11}u$$

$$v_t = bv_x + c_{22}v .$$

The first is an ordinary differential equation for u along characteristic curves defined by $\frac{dx}{dt} = -a$. The second is an o.d.e. for v along characteristic curves $\frac{dx}{dt} = -b$. These are illustrated in Figure 2.1 for the case $|a| << |b|$, $a < 0$, $b < 0$.



Characteristics for u Characteristics for v

Figure 2.1

Initial values for these ordinary differential equations are provided by initial data ($t = 0$) and boundary data ($x = 0$, in this case) for the partial differential equations. It is clear that data for v are propagated at the relatively fast speed and that data for u move at the slow speed. Boundary conditions which suppress the fast part of the solution are therefore

$$\left.\begin{array}{l} u = \text{given function} \\ v = 0 \end{array}\right\} x = 0, \quad t > 0 .$$

Conditions on u and v cannot be given at the boundary $x = 1$ for the example given here.

The same boundary conditions work in the case where C is upper triangular, i.e., $c_{21} = 0$. The second component v still satisfies the equation $v_t = bv_x + c_{22}v$, and setting $v = 0$ at the boundary prevents the boundary from influencing the interior at the fast speed. In this case v appears as a forcing term in the equation for u, but this does not matter if $v = 0$.

Trouble can arise if C is not upper triangular. In this case u appears as a forcing function in the ordinary differential equations for v along the characteristic curves $\frac{dx}{dt} = -b$. Since u is in general nonzero, v will be nonzero in the interior even if it is set equal to zero at the boundary. The boundary data will influence the solution in the interior at the fast speed by first influencing u, which in turn forces v. The boundary conditions mentioned above will of course have some desirable effects, but it would be better to have more refined boundary conditions which are more effective at reducing

the magnitude of what propagates in from the boundary at the fast speed.

Such boundary conditions can be obtained by transforming the system so as to reduce the coupling found in the lower order term. This can be thought of as a process of identifying more precisely the quantity which moves slowly and the quantity which moves rapidly. Refined boundary conditions can be obtained by setting the new fast variable equal to zero whenever permissible and then expressing this condition in terms of the original unknowns u and v.

It would suffice to transform C to lower triangular form. However, it happens that when one uses the transformation method outlined in the next section, it is almost as easy to obtain diagonal form as it is to obtain triangular form. Diagonal form seems a bit tidier, so that is what we will seek.

2.2 Outline of a Method for Uncoupling Systems of Equations

The uncoupling method used here is a method used by Taylor [10] to reduce the coupling in systems of pseudo differential equations for which the leading symbol is in block diagonal form. It is essentially a simple perturbation argument which is disguised by the language of pseudo differential operators. It is related to some uncoupling methods used by Kreiss [4] and O'Malley and Anderson [8]. We will first outline the technique using Fourier transforms in a formal way. In later sections we will make this process rigorous.

The system is $w_t = Aw_x + Cw$. We want to use Fourier transforms to reduce it to a system of ordinary differential equations and thus make it easier to analyze. We will not transform in x because this would require information about the solution outside of the boundary. That would not be appropriate in a discussion of boundary conditions. Instead, we use Fourier transforms in t. The use of such transforms will be justified later in a localization argument which uses properties of pseudo differential operators. These operators will also provide a way of handling equations with variable coefficients. Certain properties of these operators are summarized in the Appendix.

Write the system (2.1) as

$$w_x = A^{-1}w_t - A^{-1}Cw . \tag{2.2}$$

In terms of components this is

$$\frac{\partial}{\partial x}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a^{-1} & \\ & b^{-1} \end{pmatrix} \frac{\partial}{\partial t}\begin{pmatrix} u \\ v \end{pmatrix} - \begin{pmatrix} a^{-1}c_{11} & a^{-1}c_{12} \\ b^{-1}c_{21} & b^{-1}c_{22} \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}.$$

Introduce formally the Fourier transform in t. Let ξ be the dual variable, and let $\hat{u}$, $\hat{v}$, $\hat{w}$ be the transforms of u, v, w. If the coefficients of the equation are constant, (2.2) becomes

$$\begin{aligned} \hat{w}_x(x,\xi) &= i\xi A^{-1}\hat{w} - A^{-1}C\hat{w} \\ &= i\xi\left(A^{-1} - \frac{1}{i\xi}A^{-1}C\right)\hat{w} \\ &= i\xi R(i\xi)\hat{w}. \end{aligned} \tag{2.3}$$

When ξ is large the matrix $R(i\xi)$ is a perturbation of the diagonal matrix A^{-1}. We will use a perturbation argument to reduce the coupling caused by the off-diagonal elements.

Let $Q(i\xi) = I + (i\xi)^{-1}M$, where M is a matrix to be determined. For large ξ, Q^{-1} exists and has the expansion

$$Q^{-1} = I - \frac{1}{i\xi}M + \mathcal{O}\left(\frac{1}{\xi^2}\right).$$

Using (2.3), we have

$$\begin{aligned} QRQ^{-1} &= \left(I + \frac{1}{i\xi}M\right)\left(A^{-1} - \frac{1}{i\xi}A^{-1}C\right)\left(I - \frac{1}{i\xi}M + \mathcal{O}(\xi^{-2})\right) \\ &= A^{-1} + \frac{1}{i\xi}\left(MA^{-1} - A^{-1}M - A^{-1}C\right) + \mathcal{O}(\xi^{-2}) \end{aligned} \tag{2.4}$$

The coupling is of order -2 if $MA^{-1} - A^{-1}M - A^{-1}C$ is diagonal, i.e., if

$$(2.5)\quad \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}\begin{pmatrix} a^{-1} & \\ & b^{-1} \end{pmatrix} - \begin{pmatrix} a^{-1} & \\ & b^{-1} \end{pmatrix}\begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} - \begin{pmatrix} a^{-1}c_{11} & a^{-1}c_{12} \\ b^{-1}c_{21} & b^{-1}c_{22} \end{pmatrix}$$

is a diagonal matrix. We therefore set to zero the off-diagonal entries in (2.5).

$$(2.6)\quad \begin{aligned} m_{12}b^{-1} - a^{-1}m_{12} - a^{-1}c_{12} &= 0 \qquad \text{(row 1, column 2)} \\ m_{21}a^{-1} - b^{-1}m_{21} - b^{-1}c_{21} &= 0 \qquad \text{(row 2, column 1)} \end{aligned}$$

The equation (2.6) can be solved for m_{12} and m_{21} provided $a \neq b$, which is certainly true in this case.

$$m_{12} = \frac{a^{-1}c_{12}}{b^{-1} - a^{-1}} = \frac{bc_{12}}{a - b}$$

$$m_{21} = \frac{b^{-1}c_{21}}{a^{-1} - b^{-1}} = \frac{ac_{21}}{b - a}\ .$$

The entries m_{11} and m_{22} appear only in the diagonal elements of (2.6), and in fact the terms involving them cancel. These values can therefore be chosen arbitrarily. For convenience, we will take $m_{11} = 0$ and $m_{22} = 0$. The matrix Q is then given by $Q = I + (i\xi)^{-1}M$, or

$$(2.7)\quad Q = I + \frac{1}{i\xi}\begin{pmatrix} 0 & \dfrac{bc_{12}}{a - b} \\ \dfrac{ac_{21}}{b - a} & 0 \end{pmatrix}$$

Equation (2.4) then becomes

$$QRQ^{-1} = Q(A^{-1} - (i\xi)^{-1}A^{-1}C)Q^{-1}$$

$$= A^{-1} + \frac{1}{i\xi}\begin{pmatrix} -a^{-1}c_{11} & 0 \\ 0 & -b^{-1}c_{22} \end{pmatrix} + \mathcal{O}\left(\frac{1}{\xi^2}\right). \tag{2.8}$$

For later reference we estimate the coefficient of the error term $\mathcal{O}(\xi^{-2})$. A calculation based on (2.4) shows that this coefficient is a matrix whose entries are bounded by the corresponding entries of

$$\text{constant} \cdot \gamma^2 \begin{pmatrix} |b|^{-1} & |a|^{-1} \\ |b|^{-1} & |b|^{-1} \end{pmatrix}. \tag{2.9}$$

Here $\gamma = \max\{|c_{ij}|\}$, and the constant which appears first depends only on the number of terms involved and the error made in approximating $|b-a|^{-1}$ by $|b|^{-1}$ (recall $|a| << |b|$). In this case the constant is a little larger than 4.

We use (2.8) to reduce the coupling in equation (2.3).

$$\frac{\partial \hat{w}}{\partial x}(x,\xi) = (i\xi A^{-1} - A^{-1}c)\hat{w} \tag{2.3}$$

$$\frac{\partial}{\partial x}(Q\hat{w}) = Q(i\xi A^{-1} - A^{-1}c)Q^{-1}(Q\hat{w}) . \tag{2.10}$$

The entries in the matrix Q are independent of x since for this simplified treatment we have assumed that the differential equation has constant coefficients. Let $\hat{w}_1 = Qw$, and use the properties of

the similarity transformation in (2.8). Equation (2.10) becomes

$$(2.11)\quad \frac{\partial \hat{w}_1}{\partial x} = \left\{ i\xi \begin{pmatrix} a^{-1} & \\ & b^{-1} \end{pmatrix} - \begin{pmatrix} a^{-1}c_{11} & \\ & b^{-1}c_{22} \end{pmatrix} + \mathcal{O}\left(\frac{1}{\xi}\right) \right\} w .$$

When ξ is large, the coupling caused by the lower order terms in (2.11) is weaker than the coupling in the original system (2.4). This means that for large ξ we can identify more precisely the rapidly moving part of the solution and do a better job of suppressing it. This restriction to high frequencies is not a serious one, since the goal of this work is to suppress the effect of numerical errors, which are mainly high frequency phenomena. In Section 2.6 we will discuss the range of values of ξ for which the method is effective.

This method can be applied repeatedly to reduce even further the coupling at high frequencies. To reduce the coupling to $\mathcal{O}(\xi^{-2})$, we would multiply equation (2.11) by a matrix of the form $I + (i\xi)^{-2}M_2$, and then determine M_2 in the same way that we found the matrix M above. In general, to reduce the coupling from $\mathcal{O}(\xi^{-(n-1)})$ to $\mathcal{O}(\xi^{-n})$, we would use a multiplier of the form $I + \xi^{-n}M_n$. The details of this process involve no new ideas and will not be given here.

The method has been presented for 2×2 matrices. In [4], [8], [10] it is used for block matrices having two square blocks on the diagonal. In this more general case the equations corresponding to (2.6) can be solved provided that the diagonal blocks corresponding to a^{-1} and b^{-1} have disjoint spectra. This method is also valid for block matrices having any number of diagonal blocks. A general form of this method will be discussed in Section 3.3.

We now find boundary conditions for this constant coefficient system which suppress the fast part of the solution at the boundary. The new dependent variable is defined by $\hat{w}_1 = O\hat{w}$. By (2.7) this is

$$\begin{pmatrix} \hat{u}_1(x,\xi) \\ \hat{v}_1(x,\xi) \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{i\xi}\left(\frac{bc_{12}}{a-b}\right) \\ \frac{1}{i\xi}\left(\frac{ac_{21}}{b-a}\right) & 1 \end{pmatrix} \begin{pmatrix} \hat{u}(x,\xi) \\ \hat{v}(x,\xi) \end{pmatrix}$$

v_1 is our new notion of what constitutes the rapidly moving part of the solution. For large ξ it is a perturbation of the fast characteristic variable v. To suppress the fast part of the solution we set $v_1 = 0$ at $x = 0$, i.e.,

$$\frac{1}{i\xi}\left(\frac{ac_{21}}{b-a}\right)\hat{u} + \hat{v} = 0 \quad \text{at} \quad x = 0 . \tag{2.12}$$

To obtain local boundary conditions we multiply (2.12) by $i\xi$ and then apply an inverse Fourier transform. The result is

$$\frac{\partial v}{\partial t} + \left(\frac{ac_{21}}{b-a}\right)u = 0 \quad \text{at} \quad x = 0 .$$

With this we conclude the outline of the uncoupling method. There are several things left to do for problems in one space dimension. We still need to justify the use of Fourier transforms in time, present the uncoupling method for systems having variable coefficients, and discuss further the effect of the uncoupling on the behavior of the solution.

2.3 Transformations in Time

Here we discuss the question of taking Fourier transforms in time. We first recall that it is a good idea to use transforms in one variable or the other since this can simplify the analysis of the problem. Partial differential equations can be reduced to ordinary differential equations, and through these transforms the solutions can be expressed as superpositions of plane waves. This latter point is particularly important for problems in several space dimensions since in that case the direction of propagation can be as important as speed.

However, there are certain difficulties associated with the use of Fourier transforms in this case. First of all, we cannot use Fourier transforms in x since these involve information about the solution outside the boundary. This is not appropriate in a discussion of boundary conditions. On the other hand, we cannot use Fourier transforms in t directly, either. The reason is that in general the solution to a linear hyperbolic system can grow exponentially as $t \to +\infty$. This makes it impossible to define a Fourier transform either in the classical integral sense or in the sense of tempered distributions.

A common cure for this problem of exponential growth is the use of the Laplace transform. Let $s = \eta + i\xi$, where η and ξ are real and $\eta > 0$. The Laplace transform of a function $w = w(t)$ is defined by

$$Lw(s) = \int_0^\infty e^{-st}\, w(t)dt .$$

This is certainly well defined provided η is sufficiently large. However, we are reluctant to use this transform for this problem because of the effect it has on the form of the differential equation. Derivatives are transformed according to the relation

$$(Lw_t)(s) = \int_0^\infty e^{-st}\, w_t(t)dt$$

$$= sLw - w(0) .$$

The transform of the equation $w_t = Aw_x + Cw$ is therefore

$$s\hat{w}(x,s) - w(x,0) = A\hat{w}_x(x,s) + C\hat{w} , \tag{2.13}$$

where $\hat{w}(x,s)$ is the Laplace transform in t for fixed x. The trouble with (2.13) is that it includes initial values of the solution. We would like to use our transformed equation to find boundary conditions having certain properties, but the presence of the initial data in (2.13) appears to complicate matters.

These problems with the Fourier and Laplace transforms can be avoided by using certain properties of pseudo differential operators. We were going to introduce these operators anyway in order to treat systems with variable coefficients, so it is no extra trouble to use this approach to solve the transformation problem. The main idea is

to localize the solution in time in order to make the Fourier transformation possible and then show that this localization does not have a great effect on the equation.

We first recall the process outlined in the previous section. There we formally applied a Fourier transformation to the differential equation and then manipulated the transformed equation. These manipulations consisted of multiplying Fourier transforms by certain functions of the dual variables. In effect, we were applying pseudo differential operators to both sides of the differential equation. The utility of these manipulations suffered from the fact that the Fourier transformations were not justified and from the restriction to systems with constant coefficients. However, these problems disappear if we apply general pseudo differential operators directly to the given differential equation rather than first trying to find a suitable transformed equation. There will be no problem with variable coefficients, and the Fourier transformation can be treated in the manner described below.

Restrict attention to a fixed time interval $a \leq t \leq b$, and choose $\psi \in C_0^\infty(\mathbb{R})$ so that $\psi(t) = 1$ if $a \leq t \leq b$. Consider the differential equation $w_t = Aw_x + Cw$. We multiply the solution w by the cutoff function ψ to produce a function which has compact support in t and which therefore has a Fourier transform. If ψw satisfied the differential equation, then in the case of constant coefficients we could immediately apply the transformation to the differential equation. But this is not the case, since all we can say is

$$\frac{\partial}{\partial t}(\psi w) = A\frac{\partial}{\partial x}(\psi w) + C(\psi w) \tag{2.14}$$

provided $a \leq t \leq b$. Equation (2.14) holds for t in this interval because $\psi = 1$ there, but it may fail to hold for $t \notin [a,b]$.

We will not try to manipulate a transformed equation, but instead we will apply certain pseudo differential operators directly to the given differential equation. In the next section we will construct operators which uncouple the equation in a manner analogous to that described in the preceding section. The manner in which these operators will be applied requires some explanation.

Write (2.14) as

$$(\psi w)_x = A^{-1}(\psi w)_t - A^{-1}C(\psi w)\,. \tag{2.15}$$

Denote the left and right sides of (2.15) by L and R, respectively, and let P be a pseudo differential operator in t which we would apply to (2.15) in an attempt to uncouple the system. Since L and R both have compact support in t, there is no problem in forming $P(L)$ and $P(R)$. The question is whether the two are in some sense equal.

We know that $L = R$ if $a \leq t \leq b$ and $L \neq R$ for certain other t. Since pseudo differential operators are nonlocal, we can conclude that $P(L)$ and $P(R)$ are nowhere equal except perhaps at a few points where equality occurs by accident. But we can still say something about $P(L) - P(R)$ in the interval $a \leq t \leq b$ where we know that L and R are equal. Pseudo differential operators have a property

known as "pseudo locality". In this case this property implies that the difference $P(L) - P(R)$ on the interval $[a,b]$ is given by an operator of order $-\infty$. Roughly speaking, this means that the difference is very small at high frequencies. On the interval $[a,b]$ we do not have equality of $P(L)$ and $P(R)$, but instead we have a near equality which is compatible with the asymptotic nature of our method which was indicated in the preceding section.

When we uncouple the system, then, we will first choose a time interval $[a,b]$ of interest and cut off the solution outside of that interval. This will make it possible to apply pseudo differential operators to each side of the "equation" (2.15). On the interval $a \leq t \leq b$ we have $P(L) = P(R)$ modulo an error of order minus infinity. This error term will be dominated by other errors arising in the uncoupling procedure. If we are interested in analyzing the solution on a different time interval, we will have to choose a different cutoff function ψ. This will alter the equations we obtain, but only by modifying the coefficients of the error terms in certain asymptotic formulas. In the rest of this paper we will assume that the solution has been cut off in time, and we will not bother to write the cutoff functions ψ explicitly.

2.4 A More Complete Treatment of the Uncoupling Method

In this section we will use pseudo differential operators to uncouple the given system of partial differential equations. The treatment given here is similar to that given in Section 2.2, but it is more complete. This treatment is valid for systems having variable coefficients, and it uses the method of Section 2.3 for obtaining Fourier transforms in time.

As before, we will consider the equation

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a & \\ & b \end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} \tag{2.16}$$

for $0 < x < 1$ and $t > 0$. We assume $|a| << |b|$, and for the sake of definiteness we will assume $a < 0$ and $b < 0$. The system can also be written in the form $w_t = Aw_x + Cw$, where $w = (u,v)^T \in \mathbb{R}^2$. The entries in A and C are functions of x and t.

In order to simplify the notation we have chosen a system having two scalar components. The method presented here works equally well for systems having several components and for systems in partitioned form.

Write the system in the form

$$w_x = A^{-1}w_t - A^{-1}Cw. \tag{2.17}$$

In terms of components, this is

$$(2.18)\qquad \frac{\partial}{\partial x}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a^{-1} & \\ & b^{-1} \end{pmatrix} \frac{\partial}{\partial t}\begin{pmatrix} u \\ v \end{pmatrix} - \begin{pmatrix} a^{-1}c_{11} & a^{-1}c_{12} \\ b^{-1}c_{21} & b^{-1}c_{22} \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}.$$

In Section 2.2 we formally applied a Fourier transformation in t and then manipulated the transformed equation. These manipulations consisted of multiplying the transformed equation by matrices of the form $I + (i\xi)^{-1}M$ and then determining a suitable M. This reduced the coupling in the equation to $\mathcal{O}(\frac{1}{\xi})$. To reduce the coupling from $\mathcal{O}(\xi^{-(n-1)})$ to $\mathcal{O}(\xi^{-n})$, we would use multipliers of the form $I + (i\xi)^{-n}M_n$.

In this section we will not try to manipulate a transformed equation, for reasons stated earlier. Instead, we will apply certain pseudo differential operators directly to the given equation (2.18). The operator which will reduce the coupling from order zero to order -1 will have a symbol of the form $I + (i\xi)^{-1}M$. More generally, the operator reducing the coupling from order $-n+1$ to order $-n$ will have a symbol of the form $I + (i\xi)^{-n}M_n$. The process given here is similar to the process of Section 2.2, since the leading order term of the composition of two pseudo differential operators is given by the product of their symbols. The difference between this treatment and the earlier one is the presence of certain lower order correction terms appearing in the formula for composition of operators.

Write equation (2.17) in the form

$$(2.19)\qquad \frac{\partial w}{\partial x} = Gw + Dw,$$

where $D = -A^{-1}C$ and G is the operator with symbol $i\xi A^{-1}$, i.e.,

$$Gw(x,t) = \int e^{i\xi t} A^{-1}(x,t)\, i\xi\, \hat{w}(x,\xi)d\xi$$

$$= A^{-1}w_t .$$

Here $\hat{w}(x,\xi)$ denotes the Fourier transform of w with respect to t for fixed x. It is understood that w is cut off in t according to the remarks of the preceding section. We will suppress this fact in our notation.

Apply a pseudo differential operator $I + K$ to (2.19), where K is an operator of order -1 which is to be determined. The symbol of K will be $(i\xi)^{-1}M$ for some matrix M depending on x and t. From (2.19) we obtain

$$\frac{\partial}{\partial x}[(I+K)w] = (I+K)G(I+K)^{-1}[(I+K)w]$$

$$+ (I+K)D(I+K)^{-1}[(I+K)w] \tag{2.20}$$

$$+ K_x w .$$

Here K_x is the operator whose symbol is obtained by differentiating the symbol of K with respect to the parameter x. $(I+K)^{-1}$ denotes a parametrix of $I + K$. It is not hard to show that $(I+K)^{-1}$ has an asymptotic expansion

$$(I+K)^{-1} \sim I - K + K^2 - K^3 + \ldots . \tag{2.21}$$

The validity of this expansion depends on the fact that the order of

K is negative.

Let $w_1 = (I+K)w$. From (2.20) and (2.21) we obtain

$$\frac{\partial w_1}{\partial x} = (I+K)G(I-K)w_1 + Dw_1$$

$$+ \text{terms of order } (-1) \text{ or less,}$$

or

$$\frac{\partial w_1}{\partial x} = Gw_1 + (KG - GK + D)w_1 \tag{2.22}$$

$$+ (\text{order } (-1))w .$$

The operator $KG - GK + D$ appearing in (2.22) has order zero. We want to choose K so that the leading symbol of this operator is diagonal, since this would imply that the coupling in the system (2.22) has order -1.

Let σ_K, σ_G denote the symbols of K and G, respectively. The composition law for pseudo differential operators implies that the symbol of $KG - GK + D$ is

$$\sigma_K\sigma_G - \sigma_G\sigma_K + D + \text{order } (-1) . \tag{2.23}$$

Let $\sigma_K = (i\xi)^{-1}M$, where M is a matrix depending on x and t which we shall determine, and recall that $\sigma_G = i\xi A^{-1}$. (See (2.19).) The expression in (2.23) then becomes

$$MA^{-1} - A^{-1}M + D + \text{order } (-1) . \tag{2.24}$$

This is the symbol of the operator $KG - GK + D$ which appears in (2.22). The symbol of the zero-order part of (2.22) is therefore $MA^{-1} - A^{-1}M + D$,

and the equation is uncoupled to order -1 if and only if

$$MA^{-1} - A^{-1}M + D = \text{a diagonal matrix}. \tag{2.25}$$

This is exactly the condition encountered in Section 2.2, equation (2.5). (Recall that we let $D = -A^{-1}C$ in (2.19).) This can be solved for M in the same way as before. From the earlier work we conclude that the symbol of the operator K is given by

$$\sigma_K = \frac{1}{i\xi} M = \frac{1}{i\xi} \begin{pmatrix} 0 & \frac{bc_{12}}{a-b} \\ \frac{ac_{21}}{b-a} & 0 \end{pmatrix} . \tag{2.26}$$

With this choice of K the coupling in equation (2.22) has order -1.

This process can be continued indefinitely in order to reduce further the coupling in the system. To reduce the coupling to order -2, we can apply an operator $I + K_2$ to equation (2.22), where K_2 has a symbol of the form $(i\xi)^{-2}M_2$. The matrix M_2 can be determined in the same way that M was determined above. In the equation for M_2 corresponding to (2.25), the matrix corresponding to D represents the error terms of order -1 in (2.22). These terms would have to be calculated explicitly when deriving (2.22). The new dependent variable for the system would have the form $w_2 = (I + K_2)w_1 = (I + K_2)(I + K_1)w$. Further uncoupling can be carried out in the same manner.

We note that the symbol in (2.26) is the same as the one obtained in Section 2.2. This is not the case for symbols which uncouple the system further. The reason for this is that these symbols are influenced

by error terms which result from the application of the composition law for pseudo differential operators during prior applications of the uncoupling method. These error terms are generally nonzero for systems having variable coefficients, and they cannot be detected by the formal treatment of constant coefficient systems which appeared in Section 2.2.

We conclude this section by mentioning a minor technical difficulty associated with a symbol of the form $(i\xi)^{-n}M_n$. Strictly speaking, such a function cannot be a symbol of a pseudo differential operator because of the singularity at $\xi = 0$. But it can be modified in a neighborhood of $\xi = 0$ to produce a smooth bounded function of ξ. Such a change will affect only very low frequencies. From now on we will always assume that such a modification has been made, and we will ignore this fact in our notation.

2.5 Boundary Conditions

We will now use the results of the preceding section to find boundary conditions which suppress the rapidly moving part of the solution. During the uncoupling process we adopted a change of dependent variables whose effect was to weaken the coupling contained in the lower order term, at least at high frequencies. The new dependent variables can be thought of as more precise descriptions of the rapidly moving and slowly moving parts of the solution. We will find our boundary conditions by attempting to set the "fast" variable equal to zero at the boundary. In general, it is not possible to do this exactly, but it can be done to an order of accuracy which is compatible with the degree of coupling which remains in the differential equation.

We will consider the system (2.22) which was obtained through one application of the uncoupling method. Systems obtained through several applications of the process can be treated in a similar manner. Equation (2.22) is

$$\frac{\partial w_1}{\partial x} = Gw_1 + (\text{diagonal term of order } 0)w_1 + (\text{order } (-1))w \tag{2.27}$$

where

$$w_1 = (I+K)w . \tag{2.28}$$

Here $w = (u,v)^T$ is the orginal dependent variable for the system. The symbol of the operator K is given by (2.26). If we let $w_1 = (u_1,v_1)^T$, then (2.26) and (2.28) give

$$(2.29) \qquad u_1(x,t) = \int e^{i\xi t} \left[\hat{u}(x,\xi) + \frac{1}{i\xi} \left(\frac{bc_{12}}{a - b} \right) \hat{v}(x,\xi) \right] d\xi$$

$$(2.30) \qquad v_1(x,t) = \int e^{i\xi t} \left[\frac{1}{i\xi} \left(\frac{ac_{21}}{b - a} \right) \hat{u}(x,\xi) + \hat{v}(x,\xi) \right] d\xi .$$

The components u_1 and v_1 are perturbations of the original components u and v. Recall that u is associated with the slow characteristics of the system and that v is associated with the fast characteristics. The new fast variable is therefore v_1, and our goal is to set this equal to zero at the boundary whenever permissible.

This can be done easily provided the system has coefficients which are independent of t. In this case the bracketed quantity in the integral in (2.30) depends only on x and ξ, and it is therefore the Fourier transform of v_1 with respect to t for fixed x. We want $v_1 = 0$ when $x = 0$. This can be accomplished by setting the transform equal to zero at $x = 0$, so we obtain

$$(2.31) \qquad \frac{1}{i\xi} \left(\frac{ac_{21}}{b - a} \right) \hat{u}(0,\xi) + \hat{v}(0,\xi) = 0 \quad \text{for all} \quad \xi .$$

To obtain a local boundary condition, we multiply by $i\xi$ and then invert the Fourier transform. The result is

$$(2.32) \qquad \frac{\partial v}{\partial t} + \left(\frac{ac_{21}}{b - a} \right) u = 0 \quad \text{when} \quad x = 0 .$$

This argument is not valid if the coefficients in the system depend on t. In this case the bracketed quantity in the integral in (2.30) depends on t as well as x and ξ. It cannot, therefore, be

the Fourier transform of anything, and it is not possible to write (2.31). However, despite the fact that the derivation given above is invalid, it is still possible to use (2.32) as a boundary condition in the case of variable coefficients.

Proposition 2.1. If u and v satisfy (2.32), then the new "fast" variable v_1 satisfies

$$v_1 = (\text{operator of order } -2)u \quad \text{at} \quad x = 0 .$$

Proof. Suppose that (2.32) holds, and apply to it the operator whose symbol is $(i\xi)^{-1}$. This gives

$$''\left(\frac{1}{i\xi}\right)'' \circ ''(i\xi)''v + ''\left(\frac{1}{i\xi}\right)''\left(\frac{ac_{21}}{b-a}\right)u = 0 .$$

Here $''\left(\frac{1}{i\xi}\right)''$ and $''(i\xi)''$ denote pseudo differential operators with symbols $(i\xi)^{-1}$ and $i\xi$, respectively. The small circle denotes composition of operators. The composition law for pseudo differential operators yields

$$v + ''\left(\frac{1}{i\xi}\right)\left(\frac{ac_{21}}{b-a}\right)''u + ''\frac{1}{i}\frac{\partial}{\partial\xi}\left(\frac{1}{i\xi}\right)\cdot\frac{\partial}{\partial t}\left(\frac{ac_{21}}{b-a}\right)''u \tag{2.33}$$

$$+ \text{order}\,(-3) = 0 .$$

The sum of the first two terms is v_1, as can be seen by comparing (2.33) and (2.30). The third term has order -2. From this the result follows immediately. We note that the term of order -2 is generally nonzero when the coefficients of the system depend on t.

The "fast" variable v_1 is therefore small at high frequencies when $x = 0$. We will see in the next section that this order of accuracy is compatible with the degree of coupling remaining in the system of differential equations when it is written in the form (2.27).

We should note that, for this particular case, there is an easy way to find a local boundary condition which sets v_1 exactly equal to zero. Set (2.30) equal to zero, and write this as

$$\frac{ac_{21}}{b-a} \int e^{i\xi t} \left(\frac{1}{i\xi}\right) \hat{u}\, d\xi + v = 0 \quad \text{at} \quad x = 0 .$$

If $ac_{21} \neq 0$, we can multiply by $(b-a)/ac_{21}$ and then differentiate with respect to t. The result is

$$u + \frac{\partial}{\partial t}\left[\left(\frac{b-a}{ac_{21}}\right) v\right] = 0 .$$

The trouble with this approach is that it does not work if the expression has more than one term of negative order. This will be the case if the system has three or more components or if we have applied the uncoupling method more than once. In general, it is necessary to use the ideas mentioned in the proof of Proposition 2.1.

We will indicate how this process works in the case where the uncoupling method is applied again to uncouple (2.27) to order -2. In this case the method of Proposition 2.1 can be used to help generate a boundary condition, not just verify its utility.

After additional uncoupling the system becomes

$$\frac{\partial w_2}{\partial x} = (\text{diagonal operator of order } 1)w_2 + (\text{order } (-2))w$$

where $w_2 = (I + K_2)(I + K_1)w$. Here K_2 is a suitably chosen operator of order -2, and $K_1 = K$. We can take the symbol of K_2 to be zero on the diagonal, as was the case for K. w_2 can be written

$$w_2 = (I + K_1 + K_2 + K_2K_1)w \ .$$

The operator K_2K_1 has order -3 and can be deleted from the expression for the new dependent variable without affecting the order of the coupling in the system. We therefore let

$$\tilde{w}_2 = (I + K_1 + K_2)w \ , \tag{2.34}$$

and the system becomes

$$\frac{\partial \tilde{w}_2}{\partial x} = (\text{operator of order } 1)\tilde{w}_2 + (\text{order } (-2))w \ .$$

If we let $\tilde{w}_2 = (u_2, v_2)^T$, then v_2 is the new "fast" variable. According to (2.34), v_2 can be written

$$v_2(x,t) = \int e^{i\xi t} \left[\hat{v}(x,\xi) + \frac{c_1(x,t)}{i\xi} \hat{u} + \frac{c_2(x,t)}{(i\xi)^2} \hat{u} \right] d\xi \ , \tag{2.35}$$

where $c_j(i\xi)^{-j}$ is the lower left element of the symbol of K_j.

We want to set $v_2 = 0$ at $x = 0$. If c_1 and c_2 were inpendent of t, then we could set to zero the bracketed factor in

the integral in (2.35). If we clear denominators and invert the Fourier transform, the result is

$$\frac{\partial^2 v}{\partial t^2} + c_1 \frac{\partial u}{\partial t} + c_2 u = 0 \quad \text{at} \quad x = 0 . \tag{2.36}$$

The method of Proposition 2.1 can be used to show that this boundary condition still has some validity when the coefficients c_1 and c_2 vary with t. However, we can obtain a better condition for the case of variable coefficients by starting with a more general form

$$\frac{\partial^2 v}{\partial t^2} + c_1 \frac{\partial u}{\partial t} + qu = 0 \quad \text{at} \quad x = 0 , \tag{2.37}$$

and then determining q in a manner which we now describe.

Proposition 2.2. If $q = c_2 - \frac{\partial c_1}{\partial t}$, then the condition (2.37) implies that the "fast" variable v_2 satisfies

$$v_2 = (\text{operator of order } -3)u \quad \text{at} \quad x = 0 .$$

If (2.36) is satisfied, i.e., $q = c_2$, then in general we only have $v_2 = \text{order}(-2)$.

Proof. Suppose (2.37) holds, and apply to this equation the pseudo differential operator whose symbol is $(i\xi)^{-2}$. This gives

$$\left(\frac{1}{(i\xi)^2}\right) \circ (i\xi)^2 v + \left(\frac{1}{(i\xi)^2}\right) \circ (i\xi c_1) u$$

$$+ \left(\frac{1}{(i\xi)^2}\right) \circ q u = 0 .$$

We now apply the composition law for pseudo differential operators and obtain

$$v + \left(\frac{c_1}{i\xi}\right) u + \left(\frac{-2}{(i\xi)^3}\left(i\xi \frac{\partial c_1}{\partial t}\right)\right) u$$

$$+ \left(\frac{q}{(i\xi)^2}\right) u + (\text{order } (-3))u = 0 .$$

This simplifies to

$$v + \left(\frac{c_1}{i\xi}\right) u + (i\xi)^{-2}\left(q - 2\frac{\partial c_1}{\partial t}\right) u = (\text{order } (-3))u .$$

If $q - 2\frac{\partial c_1}{\partial t} = c_2$, then the left side of this equation is equal to v_2. (See (2.35).) We therefore obtain $v_2 = \text{order } (-3)$. Note that if $q = c_2$, then we only have $v_2 = \text{order } (-2)$. This completes the proof. ∎

From this it should be clear how one can find boundary conditions corresponding to arbitrary orders of uncoupling. In general, when the system is uncoupled to order $-n$, the new "fast" variable can be expressed in terms of an operator of order $-(n+1)$. We note that the composition law for pseudo differential operators can play an important role in determining these boundary conditions.

We also note that this process works for systems having more than two components. The dependent variable for a partially uncoupled system has the form $w_n = (I+K_n) \cdot \ldots \cdot (I+K_1)w$, where K_j has order $-j$. The m^{th} component of w_n has the form

(2.38) $$v + \text{terms of negative order},$$

where v is the m^{th} component of w. The process outlined above can clearly be applied to (2.38).

In this section we have not yet discussed boundary conditions for the slow part of the solution. For the system (2.1) which we are considering here, it is necessary to give a value for a slow variable at $x = 0$. One possible condition is

(2.39) $$u = \text{given function}.$$

From (2.29) we can obtain another condition,

(2.40) $$\frac{\partial u}{\partial t} + \left(\frac{bc_{12}}{a - b}\right) v = \text{given function}.$$

The second condition has little practical value for the problem considered here. It requires boundary values for a derivative of u, and in a numerical computation it would be necessary to approximate this derivative from measured values of the solution. In this paper we have assumed that the available boundary data are not particularly accurate, so we cannot expect much accuracy at all from numerical differentiation. This implies that there is little point in attempting

to use the boundary condition (2.40).

We therefore propose the boundary conditions

$$\frac{\partial v}{\partial t} + \left(\frac{ac_{21}}{b-a}\right) u = 0 \tag{2.41}$$

$$u = g, \quad \text{for} \quad x = 0,$$

where g is a given function of t. The first condition is the condition (2.32) which was obtained through one application of the uncoupling method. It is equivalent to

$$v(0,t) = v(0,0) + \int_0^t \left(\frac{ac_{21}}{b-a}\right) g(\tau)d\tau .$$

The conditions (2.41) thus prescribe values for the characteristic variables at the boundary where the characteristics enter the region. The initial-boundary value problem with these conditions must therefore be well-posed. In Section 2.8 we will present the results of some numerical calculations which compare the conditions (2.41) with the simpler conditions

$$v = 0$$

$$u = \text{given function, for} \quad x = 0 .$$

2.6 Estimates of the Size of the Fast Part of the Solution: Outline and Physical Interpretation

In this section we begin to examine the effects of the boundary conditions discussed earlier. We will first estimate the size of the "fast" part of the solution using an approach which has much the same spirit as the formal treatment of the uncoupling process given in Section 2.2. We will then give a physical interpretation of this result. In the next section we will obtain an estimate based on the more rigorous uncoupling of Section 2.4. The first estimate is not rigorous because of the limitations of Section 2.2, but its derivation is basically an elementary version of the proof of the second estimate. We therefore present the first in order to help explain and motivate the other. The basic method used here is essentially the standard technique for finding energy estimates for hyperbolic partial differential equations.

According to the discussion in Section 2.2, the system $w_t = Aw_x + Cw$ can be transformed into a system having the form

$$\frac{\partial \hat{w}_n}{\partial x}(x,\xi) = i\xi A^{-1}\hat{w}_n + (\text{diagonal matrix with terms of order zero or less})\ \hat{w}_n + \mathcal{O}(\xi^{-n})\hat{w} \tag{2.42}$$

We have assumed from the beginning that A is diagonal. In (2.42) $\hat{w}_n$ is given by

$$\hat{w}_n(x,\xi) = (I + (i\xi)^{-n}M_n) \cdot \ldots \cdot (I + (i\xi)^{-1}M_1)\hat{w} ,$$

where the matrices M_k are chosen in the manner indicated in Section 2.2. Because of the hypothesis $|a| << |b|$ in (2.1), the second component of w_n is the new "fast" variable. We need to estimate the size of this component in solutions which satisfy the boundary conditions discussed earlier.

Let $w_n = (u_n,v_n)^T$. According to (2.42) the second component satisfies the equation

$$(2.43)\quad \frac{\partial \hat{v}_n}{\partial x}(x,\xi) = i\xi b^{-1}\hat{v}_n + \sum_{k=0}^{n-1} (i\xi)^{-k}\lambda_k \hat{v}_n + h_n(x,\xi) \quad \text{for} \quad x > 0 ,$$

where h_n is an error term satisfying

$$(2.44)\qquad |h_n(x,\xi)| \leq L_n|\xi|^{-n}(|\hat{u}(x,\xi)| + |\hat{v}(x,\xi)|) .$$

The functions u and v are the components of the vector w. The coefficient b^{-1} in (2.43) is taken from the expression for the matrix A in (2.1). The coefficients λ_k and L_n can be expressed in terms of the entries in A and C. This will be done later for the case $n = 1$.

We will now use the ordinary differential equation (2.43) to estimate the size of $|\hat{v}_n|$.

Proposition 2.3. If (2.43) and (2.44) hold, then for $x \geq 0$,

$$|\hat{v}_n(x,\xi)| \le |\hat{v}_n(0,\xi)|\, e^{(R(\xi)+\sigma^2)x}$$

(2.45)
$$+ \frac{1}{\sigma\sqrt{2}} \left(\int_0^x e^{2(R+\sigma^2)(x-y)}\, |h_n(y,\xi)|^2 dy \right)^{1/2}$$

where $R(\xi) = \lambda_0 + \Sigma_{k=2}^{n-1} (i\xi)^{-k}\lambda_k$. The sum is taken for even integers k. The constant $\sigma > 0$ can be chosen arbitrarily.

Proof. In order to simplify the notation, let $q = \hat{v}_n$, $h = h_n$, and $G(\xi) = \Sigma_{k=0}^{n-1} (i\xi)^{-k}\lambda_k$. Equation (2.43) becomes

(2.46)
$$q_x = i\xi b^{-1} q + G(\xi) q + h(x,\xi) .$$

The subscript denotes differentiation.

We will find an inequality for $\frac{\partial}{\partial x} |q|^2$ *and use this to estimate* $|q|^2$. Equation (2.46) implies

$$q_x \bar{q} = i\xi b^{-1} q\bar{q} + G(\xi) q\bar{q} + h(x,\xi)\bar{q}$$

$$q\bar{q}_x = q(-i\xi b^{-1}\bar{q}) + q\overline{G(\xi)}\,\bar{q} + q\overline{h(x,\xi)} .$$

Bars denote complex conjugation. The sum of the two equations is

$$\frac{\partial}{\partial x} |q|^2 = q\bar{q}_x + q_x\bar{q}$$

(2.47)
$$= 2\mathrm{Re}[G(\xi)]\, |q|^2 + 2\mathrm{Re}[\bar{h}q]$$

$$= 2R(\xi)\, |q|^2 + 2\mathrm{Re}[\bar{h}q] ,$$

where $R(\xi)$ is the quantity defined in (2.45). The last equality follows from the fact that the λ_k are real. This in turn follows from the derivation of the partially uncoupled system (2.42). This derivation is based upon expansions in powers of $i\xi$, and all of the coefficients in these expansions are real.

The second term on the right in (2.47) can be estimated using the inequality $2ab \le a^2 + b^2$. For any real $\sigma \neq 0$ we obtain

$$2\,\mathrm{Re}[\bar{h}q] \le 2|hq| \le 2\sigma^2|q|^2 + \frac{1}{2\sigma^2}|h|^2 .$$

With (2.47) this yields

$$\frac{\partial}{\partial x}|q|^2 \le 2(R(\xi) + \sigma^2)\,|q|^2 + \frac{1}{2\sigma^2}|h|^2 .$$

We now apply the Gronwall inequality. That is, we move the first term on the right over to the left side, multiply by the integrating factor $\exp[-2(R+\sigma^2)x]$, and then integrate. The result is

$$|q(x)|^2 \le e^{2(R+\sigma^2)x}\,|q(0)|^2 + \frac{1}{2\sigma^2}\int_0^x e^{2(R+\sigma^2)(x-y)}|h(y)|^2 dy .$$

To obtain the final result (2.45) we recall that $q = \hat{v}_n$ and use the fact that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ when $a,b \ge 0$. This completes the proof. ■

In the estimate (2.45) there is a term which involves the values of v_n at $x = 0$. According to the formal treatment of constant-

coefficient problems given in Section 2.2, it would be possible to set v_n exactly equal to zero at the boundary $x = 0$. In that case the term involving $\hat{v}_n(0,\xi)$ would not appear in (2.45). However, we saw in Section 2.5 that it is not quite possible to satisfy $v_n = 0$ when the coefficients vary with t. We have therefore included the extra term in (2.45) in order to suggest the more general behavior of the boundary conditions. According to Proposition 2.2 and the comment which followed, it is possible to obtain $\hat{v}_n(0,\xi) = \mathcal{O}(\xi^{-n-1})\hat{w}$. This is certainly compatible with the other term in (2.45), which according to (2.44) is $\mathcal{O}(\xi^{-n}\hat{w})$. We will see that something like this actually happens with the more rigorous estimate which we will obtain in the next section.

The method used here to uncouple the system is valid asymptotically as $\xi \to \infty$. It would be good to have estimates of the coefficients in error terms such as (2.44) in order to have a rough idea of the range of ξ for which the method works. In the following proposition we do this for (2.44) in the case $n = 1$, and we also give the value of a relevant parameter appearing in the estimate (2.45).

Proposition 2.4. The parameter λ_0 appearing in (2.45) is given by $\lambda_0 = -c_{22}b^{-1}$. When $n = 1$, the constant L_n in (2.44) can be taken to be

$$L_1 = \text{constant} \cdot \gamma^2 |b|^{-1}, \tag{2.48}$$

where $\gamma = \max\{|c_{ij}|\}$. The c_{ij} are the coefficients in the undifferentiated

term in (2.1). The "constant" is approximately equal to 4 in this case. (See (2.9) and the discussion which follows.)

Proof. In (2.43) the parameter λ_0 is defined to be the coefficient in the zero-order term in the differential equation for $\hat{v}_n$. For the case $n = 1$ this equation is given by the second row of (2.11). There we see $\lambda_0 = -c_{22}b^{-1}$. This value does not change with n, since further uncoupling is obtained through transformations involving matrices of the form $I + (i\xi)^{-k}M_k$ for $k \geq 2$. Such transformations cannot alter the term of order zero.

The parameter L_n in (2.44) is part of the bound on the error term in the partially uncoupled system (2.43). For the case $n = 1$ this system is given by (2.11). The coefficient of the error $O(\xi^{-1})$ is bounded by the matrix in (2.9). From this the conclusion can be read immediately. This completes the proof. ■

We pause to interpret this result. When $n = 1$ the "fast" variable v_n satisfies

$$\frac{\partial \hat{v}_1}{\partial x}(x,\xi) = (i\xi b^{-1} + \lambda_0)\hat{v}_1 + h_1(x,\xi), \tag{2.49}$$

where

$$|h_1(x,\xi)| \leq \text{constant} \cdot \frac{\gamma}{|\xi|} \cdot \frac{\gamma}{|b|}(|\hat{u}| + |\hat{v}|). \tag{2.50}$$

The first line is equation (2.43) for the case $n = 1$. The second is a consequence of (2.44) and (2.48). We compare this to the situation in which we do not do any uncoupling, but instead use the original

system $w_t = Aw_x + Cw$ given in (2.1). In that case the "fast" variable is taken to be v, the second component of the vector w. It satisfies

$$\frac{\partial \hat{v}}{\partial x} = b^{-1}\hat{v} + b^{-1}(c_{21}\hat{u} + c_{22}\hat{v}) .$$

The forcing term in this equation is dominated by $\gamma|b^{-1}\hat{u}|$, since $\gamma = \max\{|c_{ij}|\}$. A comparison with (2.50) shows that the uncoupling method used to obtain (2.49) has a substantial effect when $|\xi| >> \gamma$. This relation defines what we will mean by "large frequencies" in the context of this paper. We note that similar relations hold more generally and that ξ and γ both have the dimensions time^{-1}.

For large scale meteorological problems the Coriolis parameter is usually the dominant entry in the coefficient matrix of the lower order term in linearized systems. It is given by $f = 2\Omega \sin \varphi$, where Ω is the earth's angular velocity and φ is the angle of latitude. That is,

$$f = 2 \cdot \frac{2\pi}{24} \sin \varphi \ \text{hr}^{-1} \approx \frac{1}{2} \sin \varphi \ \text{hr}^{-1}$$

The methods discussed here should therefore work well for those time frequencies whose order of magnitude is roughly $1\ \text{hr}^{-1}$ or greater.

In the case of smaller scale problems the Coriolis parameter may be dominated by certain terms which arise in the linearization of the system. This will reduce the maximum wavelength for which the uncoupling method is effective. However, the size of the computation domain is also reduced, so it appears that the method may still be useful.

In the estimate (2.45) for $\hat{v}_n$ there are factors which allow exponential growth in x for $x > 0$. (The boundary of the domain is given by $x = 0$.) We wish to make a rough estimate of the coefficient in these exponents in order to determine the length scale on which this exponential growth can take place. One of these factors is $\exp[(R+\sigma^2)x]$, where

$$R(\xi) = \lambda_0 + \sum_{k=2}^{n-1} (i\xi)^{-k}\lambda_k .$$

The sum is taken for even k, and the parameter σ can be chosen arbitrarily. A similar factor appears in the integral term in (2.45). An analysis of the uncoupling process shows that the dependence on k of $|\lambda_k|$ is given roughly by γ^{k+1}. We omit the details of this, but the main idea is that the expansions appearing in Section 2.2 are dominated by expansions in $\gamma|\xi|^{-1}$. The behavior of $R(\xi)$ for large ξ is therefore governed by the leading term λ_0, which according to Proposition 2.4 is equal to $-c_{22}b^{-1}$. This satisfies

$$|\lambda_0| \leq \frac{\gamma}{|b|} = \frac{1}{|b|\gamma^{-1}} . \tag{2.51}$$

We recall that $\xi = \gamma$ is approximately the lowest time frequency for which the uncoupling method can have an effect. This corresponds to a period of γ^{-1}. The denominator $|b|\gamma^{-1}$ in (2.51) is therefore a rough approximation to the length of the longest fast wave for which the method applies. This defines the length scale on which the exponential growth can take place, since for large ξ the parameter λ_0 dominates the coefficient in the exponential factor $\exp[(R+\sigma^2)x]$.

2.7 Estimates of the Size of the Fast Part of the Solution: A More Complete Treatment

We turn now to the problem of making our estimates more rigorous. The estimate (2.45) for the fast part of the solution was obtained by considering an ordinary differential equation in x for the Fourier transform in time. The equation was obtained through the uncoupling process of Section 2.2. This approach gives a rough idea of how the boundary conditions affect the solution, but the result cannot be considered very rigorous. First of all, it obviously cannot work when the coefficients in the system vary with t. Furthermore, this approach ignores the problems mentioned earlier regarding Fourier transforms in time. A correct uncoupling of the system really must be based on the use of pseudo differential operators, even if the system has constant coefficients, and a correct analysis of the effect of the boundary conditions must be based on this uncoupling. In this section we give such an analysis.

As before, we will obtain estimates which indicate the behavior as $\xi \to \infty$ of the Fourier transform of the "fast" dependent variable in the system. In the earlier case we did this by estimating $\hat{v}_n(x,\xi)$ for each fixed ξ. However, in a truly rigorous treatment it is not possible to analyze this problem one frequency at a time. Instead, we will obtain analogous results by estimating Sobolev norms of the fast part of the solution. For any real number s, the norm in the Sobolev space H^s is defined by

$$\|u\|_s = \left(\int (1 + |\xi|^2)^s \, |\hat{u}(\xi)|^2 \, d\xi \right)^{1/2}$$

$$= \| \Lambda^s u \|_{L^2}$$

where

$$(\Lambda^s u)^{\wedge}(\xi) = (1 + |\xi|^2)^{s/2} \, \hat{u}(\xi) . \tag{2.52}$$

An estimate involving a Sobolev norm makes a statement about the behavior of the Fourier transform as $\xi \to \infty$. From certain such estimates we will be able to conclude that the "fast" dependent variable is small at large frequencies. However, we will not be able to conclude that the variable is small altogether, since the estimates will not say anything about low frequencies. But this is not a severe loss. The uncoupling process has an effect only at large frequencies, so it is only at such frequencies that we can identify the fast and slow parts of the solution. At low frequencies we do not know whether the "fast" variable is small, but on the other hand we do not know that it is really "fast", either.

The estimates will be obtained through a technique which resembles the one used earlier to obtain (2.45). It is essentially the standard technique for proving energy estimates for hyperbolic partial differential equations.

According to the discussion in Section 2.4, the system $w_t = Aw_x + Cw$ can be transformed into a system having the form

$$\frac{\partial w_n}{\partial x}(x,t) = Gw_n + \Sigma_n w_n + E_n w . \tag{2.53}$$

Here G is the operator with symbol

$$i\xi \begin{pmatrix} a^{-1} & \\ & b^{-1} \end{pmatrix},$$

where a and b are defined in (2.1). Σ_n is a pseudo differential operator in the time variable. It has order zero, and its symbol is a diagonal matrix in which x may appear as a parameter. E_n is an operator of order $-n$ which does not in general have a diagonal symbol and which therefore represents the error in the uncoupling process.

In (2.53) we should also include an error term which represents the effect of the procedure given in Section 2.3 for justifying the use of Fourier transforms in time. However, this term can be neglected according to a localization argument which we will present a little later. We will first derive the estimates.

Proposition 2.5. Suppose that (2.53) holds, and let $w_n = (u_n, v_n)^T$. Then for any real s there exist constants c_1 and c_2 such that for $x \geq 0$ the "fast" variable v_n satisfies

$$\begin{aligned} \|v_n(x,\cdot)\|_s^2 &\leq e^{c_1 x} \|v_n(0,\cdot)\|_s^2 \\ &\quad + c_2 \int_0^x e^{c_1(x-y)} \|w(y,\cdot)\|_{s-n}^2 \, dy, \end{aligned} \tag{2.54}$$

provided that all of the norms are finite. The norms are Sobolev norms in t for fixed x. The constants c_1 and c_2 may depend on n and s.

Proof. According to (2.53) the component v_n satisfies the equation

$$\frac{\partial v_n}{\partial x}(x,t) = L_n v_n + R_n w ,$$

where R_n is a pseudo differential operator of order $-n$, and L_n is an operator of order one with leading symbol $i\xi b^{-1}$. The symbol of R_n is a 1×2 matrix. In order to simplify the notation we let $q = v_n$, delete the subscripts in L_n and R_n, and use a subscript to denote differentiation. The result is

$$q_x(x,t) = Lq + Rw . \tag{2.55}$$

We will obtain an inequality for $\frac{\partial}{\partial x}\|q(x,\cdot)\|_s^2$ and then use this to estimate $\|q\|_s^2$. This norm is given by

$$\|q\|_s^2 = \|\Lambda^s q\|_{L^2}^2 = (\Lambda^s q, \Lambda^s q) ,$$

where Λ^s is the operator defined in (2.52). This is an operator in t which commutes with differentiation with respect to x. Therefore

$$\frac{\partial}{\partial x}\|q(x,\cdot)\|_s^2 = (\Lambda^s q, \Lambda^s q_x) + (\Lambda^s q_x, \Lambda^s q) . \tag{2.56}$$

This can be evaluated using (2.55).

$$\begin{aligned} \Lambda^s q_x &= \Lambda^s Lq + \Lambda^s Rw \\ &= L(\Lambda^s q) + [\Lambda^s, L]q + \Lambda^s Rw . \end{aligned} \tag{2.57}$$

Here $[\Lambda^s, L]$ denotes the commutator $\Lambda^s L - L\Lambda^s$. We insert (2.57) into (2.56) and obtain

$$\begin{aligned}\frac{\partial}{\partial x}\|q\|_s^2 &= (\Lambda^s q, L\Lambda^s q) + (L\Lambda^s q, \Lambda^s q) \\ &+ (\Lambda^s q, [\Lambda^s, L]q) + ([\Lambda^s, L]q, \Lambda^s q) \\ &+ (\Lambda^s q, \Lambda^s Rw) + (\Lambda^s Rw, \Lambda^s q) .\end{aligned} \tag{2.58}$$

The terms in (2.58) will be estimated using the general fact that every pseudo differential operator of order m is a bounded linear mapping from H^t into H^{t-m}.

The first row in (2.58) is equal to $((L+L^*)\Lambda^s q, \Lambda^s q)$, where L^* is the adjoint of the operator L. According to the Schwarz inequality this is bounded by

$$\|(L+L^*)\Lambda^s q\|_{L^2}\|\Lambda^s q\|_{L^2} . \tag{2.59}$$

The first factor can be estimated by observing that $L + L^*$ has order zero, even though L and L^* each have order one. This is a consequence of the fact that the leading symbol of the adjoint operator is equal to the adjoint of the leading symbol of the original operator. Since $L + L^*$ is therefore a bounded operator on L^2, it follows that (2.59) can be bounded by a constant multiple of $\|\Lambda^s q\|_{L^2}^2$, or $\|q\|_s^2$.

The second row in (2.58) involves the commutator $[\Lambda^s, L] = \Lambda^s L - L\Lambda^s$. This operator has order s, since the leading symbol of the product of two operators is given by the product of their leading symbols. The commutator is therefore a bounded mapping from H^s into

H^0 (or L^2). It follows from this and the Schwarz inequality that the second row of (2.58) can be bounded by a constant multiple of $\|q\|_s^2$.

The third row in (2.58) is dominated by a multiple of $\|q\|_s\|w\|_{s-n}$, since the operator $\Lambda^s R$ has order $s-n$. From (2.58) we can therefor conclude

$$\frac{\partial}{\partial x}\|q(x,\cdot)\|_s^2 \le K_1\|q\|_s^2 + K_2\|q\|_s\|w\|_{s-n},$$

for suitable constants K_1 and K_2. This inequality can be integrated in the same manner as a similar inequality which appeared in the proof of Proposition 3.2. The result is (2.54). This completes the proof. ■

The estimate (2.54) expresses the smoothness of the "fast" part v_n in terms of the smoothness of its boundary values $v_n(0,\cdot)$ and the smoothness of the entire solution $w = (u,v)^T$. According to the comments of Section 2.5 it is possible to choose boundary conditions for this one-dimensional case so that $v_n(0,\cdot)$ is given by an operator of order $-(n+1)$ acting on $w(0,\cdot)$. The inequality (2.54) can therefor be written

$$\|v_n(x,\cdot)\|_s^2 \le c_3 e^{c_1 x}\|w(0,\cdot)\|_{s-n-1}^2 + c_2\int_0^x e^{c_1(x-y)}\|w(y,\cdot)\|_{s-n-1}^2$$

If $w(y,\cdot)$ is in H^r for each y, then we can let $s-n=r$ and conclude that $v_n(x,\cdot)$ is in H^{n+r} for each x. The fast part of the solution is therefore n degrees smoother than the full solution.

This argument is circular as presented, since the derivation of the estimate (2.54) is based on the assumption that all of the norms are finite. This is not a real difficulty. Suppose that we have a solution w to (2.1) which satisfies the special boundary conditions and which lies in H^r for each x. We need to show that for each x the fast part v_n lies in H^{r+n}. We first note that the equation (2.55), $q_x = Lq + Rw$, has a solution in H^{r+n} which is equal to v_n when $x = 0$. This follows from a little functional analysis and the a priori estimate just obtained. See, for example, pp. 63-65 in Taylor [9]. This function q must in fact be equal to v_n for all x, since v_n is in H^s for sufficiently low s, and for such s the estimate (2.54) implies uniqueness of solutions of equation (2.55). We can conclude that the fast part has the smoothness properties desired.

One matter which we still need to consider is the error term mentioned earlier which should have appeared in (2.53). This term represents the effect of the procedure introduced in Section 2.3 for justifying the use of Fourier transforms in time. The main idea of this method is to choose a time interval $[a,b]$ of interest and then truncate the solution outside that interval by multiplying it by a smooth function with compact support. This introduces an error term in the differential equation which, on the interval $[a,b]$, can be represented by a smoothing operator. Outside $[a,b]$ it cannot in general be represented in such a manner. This error term really should appear in (2.53), but it is possible to omit this term if we localize the solution so that the behavior outside $[a,b]$ becomes

irrelevant. We describe how to do this now.

The procedure is based on a construction used by Hörmander in the proof of a theorem on propagation of singularities for linear partial differential equations. See Hörmander [3] or Nirenberg [6], p. 44. We will consider an equation $Pu = f$, where P is a pseudo differential operator. In our application this equation is (2.53) with the extra error term added, and P is an operator in both x and t. Hörmander describes a method for localizing the solution to a neighborhood of a given bicharacteristic of P. He constructs a pseudo differential operator B of order zero so that the commutator $[P,B] = PB - BP$ has order $-\infty$ and so that the symbol of B vanishes outside a conical neighborhood of the bicharacteristic. The equation $Pu = f$ implies $P(Bu) = BPu + [P,B]u$, or

$$P(Bu) = Bf + [P,B]u . \tag{2.60}$$

The local smoothness of u is given by the global smoothness of Bu (see the next paragraph), so it is possible to study the smoothness of u along the bicharacteristic by considering global estimates for (2.60). This behavior is not influenced by the term $[P,B]u$ because this term is automatically smooth in both x and t. It is also not influenced by the values of f away from the bicharacteristic, since these values are cut off by the operator B.

The fact that the local smoothness of u is determined by the global smoothness of Bu is a consequence of the fact that B is an elliptic operator of order zero in a neighborhood of the bicharacteristic. To show this rigorously one needs to do a certain amount of

work with cutoff functions. The necessary arguments are given in part (b) of the proof of Lemma 3 on page 42 of Nirenberg [6].

This localization process enables us to handle the extra error term mentioned earlier. Suppose that the symbol of B is truncated so that it is zero after the characteristic leaves the time interval $[a,b]$. Re-write equation (2.53) as

$$q_x(x,t) = Lq + Rw + Ew ,$$

where Ew is the extra error term. When we apply the operator B to this equation, this term is replaced by BEw. The support of B in time is contained in the interval $[a,b]$, and the singular behavior of Ew is confined to the complement of $[a,b]$. These facts, together with the pseudo local property of pseudo differential operators, imply that BEw must be entirely smooth in t. This term can therefore be treated as a forcing term which lies in any Sobolev class we desire. It follows that the estimate in Proposition 2.5 is completely valid provided that v_n is replaced by Bv_n and a suitable norm of BEw is inserted. The conclusions about smoothness can then be applied to Bv_n.

The method used here actually gives more precise information than is implied by Proposition 2.5, since it deals with propagation along individual bicharacteristics. This feature will be useful in the study of problems in several space dimensions, where the direction of propagation can play a key role. In particular, it will be more

important to suppress fast waves moving in a direction normal to the boundary than it will be to suppress waves moving in a nearly tangential direction. The method given here allows one to distinguish between these directions.

2.8 Numerical Computations

In this section we present the results of some numerical computations involving the boundary conditions which were derived earlier. We consider the system

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -1 & \\ & -5 \end{pmatrix} \frac{\partial}{\partial x}\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 10 & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} \tag{2.61}$$

for $0 < x < 1$ and $t > 0$. This is the system (2.1), where $a = -1$, $b = -5$, $c_{21} = 10$, and the other c_{ij} are zero. We compare the boundary conditions

$$\begin{aligned} v &= 0 \\ u &= g \qquad\qquad (x = 0) \end{aligned} \tag{2.62}$$

and

$$\begin{aligned} &\frac{\partial v}{\partial t} + \left(\frac{ac_{21}}{b-a}\right) u = 0 \\ &u = g \qquad\qquad (x = 0)\,. \end{aligned} \tag{2.63}$$

Here g is a given function of t. The first condition in (2.63) is the condition (2.32) which was obtained from the results of the uncoupling process.

In our computation the system is approximated by the leap frog difference scheme. The function g in the boundary conditions is equal to a half period of a sine wave which is extended by zero. A forward difference is used to approximate the derivative in (2.63). The surfaces pictured in Figures 2.3 and 2.4 are graphs of $\sqrt{u^2 + v^2}$ as

a function of x and t. In Figure 2.2 we illustrate the configuration of these surface plots.

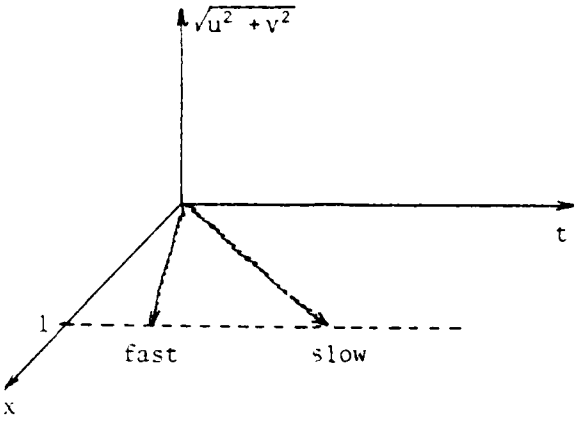


Figure 2.2

In the computations we set the solution equal to zero when t = 0. The nonzero part of the solution is due entirely to the nonzero boundary data, so it is possible to study the influence of the boundary data by examining the size of the solution in various parts of the (x,t) plane. The solution corresponding to the simple boundary condition (2.62) is graphed in Figure 2.3, and the solution corresponding to the more refined condition (2.63) is given in Figure 2.4. It is clear from the figures that the second condition is much more effective in suppressing the fast part of the solution.

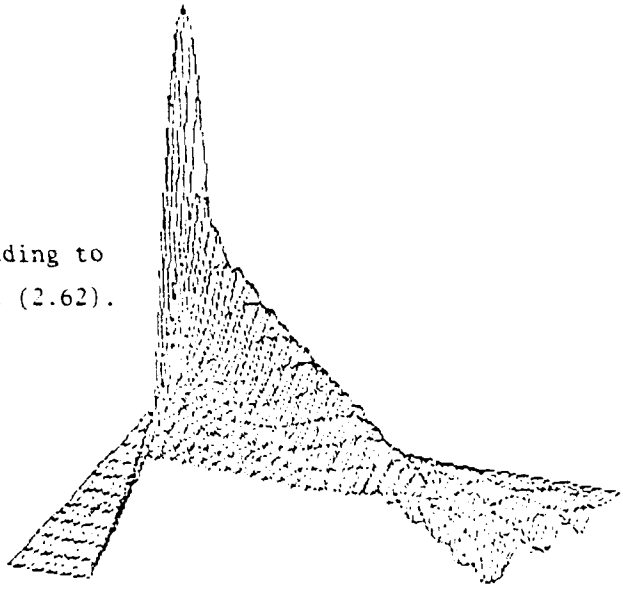


Figure 2.3. Solution corresponding to boundary condition (2.62).

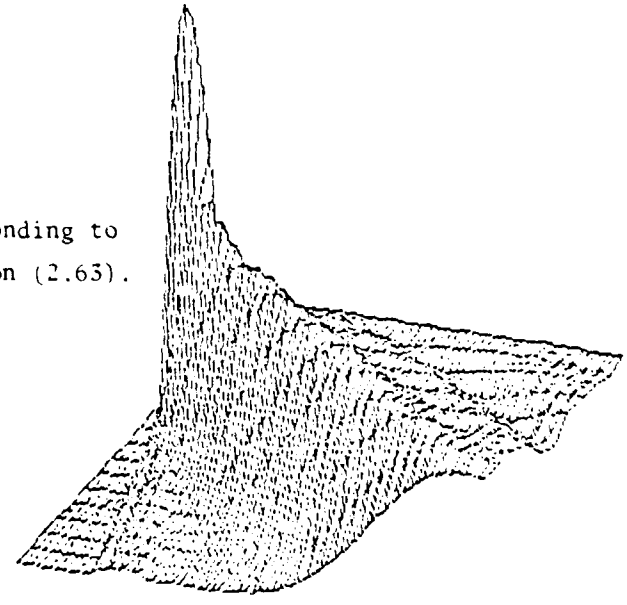


Figure 2.4. Solution corresponding to boundary condition (2.63).

CHAPTER 3

THE PROBLEM IN SEVERAL SPACE DIMENSIONS

In this chapter we will generalize the methods of the preceding chapter to systems in more than one space dimension. The major change required lies in the process of diagonalizing the leading symbol of the differential equation. In the case of one space dimension this diagonalization causes no trouble at all, but in the case of several space dimensions it can get rather involved. Otherwise, there is little difference between this case and the one discussed earlier. The lower order term can be uncoupled in exactly the same manner as before, and the fast part of the solution can still be estimated by localizing the solution to a bicharacteristic and then applying energy estimates.

We will first discuss the problem of diagonalizing the leading symbol. We will then summarize the uncoupling process for the multi-dimensional case, and we will conclude with a useful perturbation lemma which is a generalization of a technique that was used earlier. In the next chapter we will apply these methods to the shallow water equations.

3.1 Properties of the Principal Symbol

We will consider the hyperbolic system

$$w_t = Aw_x + Bw_y + Cw \tag{3.1}$$

for $x > 0$, $y \in \mathbb{R}$. Here $w(x,y,t) \in \mathbb{R}^n$, and A, B, and C are real $n \times n$ matrices which are functions of x, y, and t. Without loss

of generality we will assume that A is diagonal. In order to simplify the notation we have chosen a system in two space dimensions. Throughout this chapter it will be obvious that the discussion is equally valid for systems in higher dimensions where $x > 0$ and $y \in \mathbb{R}^k$ for $k \geq 2$.

There is no serious loss of generality in assuming that the spatial domain is a half-space. If the given domain does not have this form but still has a smooth boundary, then it is possible to localize the problem with a partition of unity and then map each boundary portion into the boundary of a half-space. In the new coordinates the problem will have the form given above.

The system (3.1) has been assumed to be hyperbolic. In this paper this will mean that for every real ζ and ω and for every point (x,y,t), the symbol

$$\zeta A + \omega B \tag{3.2}$$

has real eigenvalues and a complete set of eigenvectors. There will be no need to assume that A and B are symmetric or that the system is strictly hyperbolic.

In order to have a system with at least two time scales, we will assume that certain eigenvalues of the symbol (3.2) are substantially greater in magnitude than the others. In the case of the linearized shallow water equations this symbol has eigenvalues $-u \cdot \sigma$ and $-u \cdot \sigma \pm |c|$, where $u = (u_1,u_2)$ is the velocity of the flow about which the system has been linearized, and σ is the vector of dual variables (ζ,ω). If $|u| << c$, then this sytem has two time scales.

There is a similar set of eigenvalues for the three-dimensional, five-component Euler system for gas dynamics. In this case the small eigenvalue has multiplicity three.

We now turn to the main problem. We wish to find boundary conditions for (3.1) which prevent rapidly moving waves from entering the given spatial domain. Our plan is to first transform the system to an approximate diagonal form, or at least block diagonal form, so that each of the new dependent variables can be identified as a slow, incoming fast, or outgoing fast portion of the solution. We will then attempt to set the incoming fast components equal to zero.

The immediate goal is to diagonalize the leading order terms in the system (3.1). It would actually suffice to obtain a block diagonal form, since there is no need to separate various incoming fast components or various slow components. After this part of the uncoupling has been accomplished, we can use the methods of the preceding chapter to reduce the coupling caused by the lower order terms.

We have assumed that the matrix A in (3.1) is already in diagonal form. This involves no loss of generality, since if A is not in that form we can find a similarity transformation which makes it diagonal and then adopt a suitable change of dependent variable. Unfortunately, it is not true in general that this transformation can also diagonalize the matrix B. It is therefore necessary to do something extra if we want to diagonalize the entire principal part of (3.1).

In the case of constant coefficients it may be tempting to use Fourier transforms in x and y. This would yield the equation

$$\hat{w}(\zeta,\omega,t) = (i\xi A + i\omega B)\hat{w} + C\hat{w} .$$

The leading symbol of this equation can be diagonalized easily because it is a scalar multiple of the symbol (3.2) discussed earlier. However, the use of Fourier transforms in x requires the use of information about the solution away from the boundary $x = 0$, and this is not appropriate in a discussion of boundary conditions. It is therefore necessary to take a different approach.

We will instead use Fourier transforms in time and in the tangent variable y. For the time being we will use these transforms in a rather formal way, and it will be understood that one can obtain rigorous results by translating various arguments into the language of pseudo differential operators. We first write the system (3.1) in the form

$$w_x = A^{-1}w_t - A^{-1}Bw_y - A^{-1}Cw . \tag{3.3}$$

It will be assumed throughout this discussion that the matrix A is invertible. Let $\hat{w}(x,\omega,\xi)$ denote the Fourier transform of w with respect to y and t for fixed x. Equation (3.3) implies

$$\hat{w}_x(x,\omega,\xi) = (i\xi A^{-1} - i\omega A^{-1}B)\hat{w} - A^{-1}C\hat{w} . \tag{3.4}$$

We need to determine the values of ω and ξ for which the symbol

$$\xi A^{-1} - \omega A^{-1}B \tag{3.5}$$

can be diagonalized, and we must determine whether such a diagonalization can produce a transformed system in which each component of the

dependent variable can be identified as slow, incoming fast, or outgoing fast. The answers to these questions are not immediately obvious, since we have chosen a nonstandard set of variables in which to apply Fourier transforms.

In order to get started we must consider the eigenvalues and eigenvectors of the symbol (3.5). Suppose that ζ is a real eigenvalue of (3.5) and that v is a corresponding eigenvector. This means that

$$(\xi A^{-1} - \omega A^{-1}B)v = \zeta v . \tag{3.6}$$

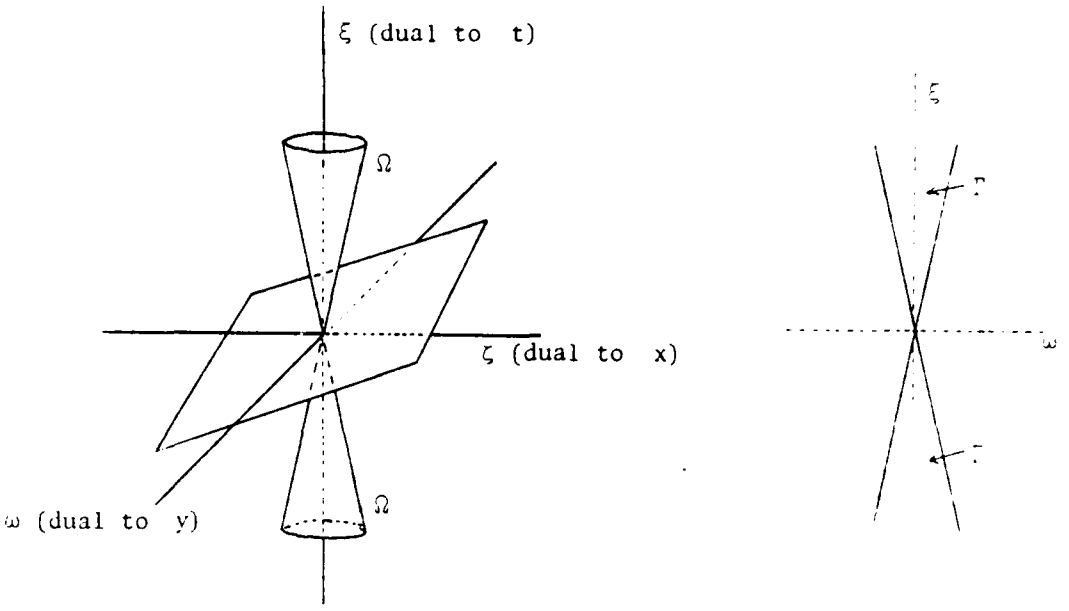
If we multiply by A and rearrange the terms, the result is

$$(\zeta A + \omega B)v = \xi v . \tag{3.7}$$

The matrix $\zeta A + \omega B$ is the symbol (3.2) which we would obtain by writing the system in the more common form (3.1) and then applying Fourier transforms in the usual variables x and y. According to (3.6) and (3.7), this symbol imposes the same relations between the dual variables ζ, ω, and ξ as the symbol (3.5), and it is possible to find the eigenvectors of one symbol by examining the eigenvectors of the other. The difference between the two situations is that in one case the variable ζ is treated as a function of ω and ξ, and in the other case ξ is treated as a function of ζ and ω. This correspondence between the two symbols will be very useful in studying (3.5). At this point in the discussion we know a great deal more about (3.2) than we do about (3.5), and the correspondence between the two will enable us to translate information about one into

information about the other.

We begin by discussing the eigenvalues of (3.2). In order to have a system with multiple time scales we have assumed that certain eigenvalues of (3.2) are considerably larger than the others. An example of such a set of eigenvalues is graphed in Figure 3.1(a). In this example there are two relatively large eigenvalues and one smaller eigenvalue for each ζ and ω. This is the configuration for the shallow water equations, and it is similar to the configuration for the Euler equations of gas dynamics. In the latter case the small eigenvalue has multiplicity three. Throughout this discussion we will assume that the largest eigenvalues of (3.2) occur in pairs and have graphs which are similar to the graphs of the large eigenvalues in Figure 3.1(a). That is, we will assume that there is a large positive eigenvalue whose graph is a narrow cone, though not necessarily a right circular cone. This implies that there must also be a negative cone, since if (ζ,ω,ξ) is a solution of (3.7) then so is $(-\zeta, \omega-, -\xi)$. These eigenvalues generate rapidly moving waves in all directions. The fact that the graphs are not necessarily right circular cones means that the speed can vary somewhat with the direction of propagation. We will denote by Ω the double cone which corresponds to the largest eigenvalues, and we will denote by Γ the portion of the (ω,ξ) space which lies inside Ω. These are labeled in Figure 3.1.



(a) Graph to the relation (3.6), (3.7)

(b) Cross-section in the (ω,ξ) space

Figure 3.1

We are now in a position to discuss the eigenvalues ζ of the symbol (3.5), $\xi A^{-1} - \omega A^{-1}B$. The quantities ζ, ω, and ξ must satisfy the relation (3.7), which is the same as the relation (3.6) which was discussed in the preceding paragraph. We can therefore study the behavior of ζ from graphs like Figure 3.1(a).

First of all, it is apparent that the number of real eigenvalues must vary with the position of (ω,ξ). If (ω,ξ) lies in Γ, then there are two values of ζ which are associated with the surface Ω. One is positive and the other is negative. As (ω,ξ) approaches the boundary of Γ, these values of ζ approach zero, and when (ω,ξ) leaves Γ the eigenvalues leave the real axis and form a pair of complex conjugates. The eigenvalues cannot be real, since for any

real ζ the point (ζ,ω,ξ) must lie on one of the surfaces in Figure 3.1(a). They are complex conjugates because they are eigenvalues of a real matrix.

It is safe to assume that the other values of ζ do not behave in this manner, at least in a neighborhood of Γ. In the case of the shallow water equations the other ζ satisfies the equation $\xi = u_1\zeta + u_2\omega$. We have assumed that the matrix A in (3.1) is non-singular, which in this case is equivalent to saying $u_1 \neq 0$. It is therefore possible to solve for ζ in terms of ω and ξ, whether or not (ω,ξ) is in Γ. A similar situation holds for the Euler equations of gas dynamics. We will therefore assume in general that for (ω,ξ) in a neighborhood of Γ there is no problem in solving for the values of ζ associated with surfaces different from Ω.

We will now characterize the behavior of (3.5) when (ω,ξ) lies in Γ. This is the only portion of the (ω,ξ) space in which we are really interested, since this is the only portion which corresponds to the rapidly moving waves. We will say more about this a little later.

Proposition 3.1. If (ω,ξ) is in Γ, then the symbol (3.5), $\xi A^{-1} - \omega A^{-1}B$, has real eigenvalues and a complete set of real eigenvectors. This is not the case if (ω,ξ) is not in Γ. The eigenvectors can be determined from those of the symbol (3.2), $\zeta A + \omega B$.

Proof. Equations (3.6) and (3.7) show that the eigenvectors of (3.2) are also eigenvectors of (3.5). We know that (3.2) has a complete set of real eigenvectors corresponding to fixed (ζ,ω) and

various eigenvalues ξ. We want to show the same thing for (3.5), for fixed (ω,ξ) in Γ and various eigenvalues ζ.

Suppose that (ω,ξ) is in Γ, and let $\zeta_1,\ldots,\zeta_m$ denote the eigenvalues of (3.5). For each ζ_j choose a basis B_j for the eigenspace of $\zeta_j A + \omega B$ corresponding to the eigenvalue ξ. We are allowing for the possibility that the symbol (3.2) might have multiple eigenvalues. The elements of B_j are also eigenvectors of (3.5) corresponding to the eigenvalue ζ_j. We claim that the union of the B_j is a complete set of vectors. There are clearly enough of these vectors. The fact that they are linearly independent follows from an argument which is essentially the one which shows that eigenvectors corresponding to distinct eigenvalues are linearly independent. This completes the proof. ■

The matters discussed in this section can be given a physical interpretation. Suppose that the coefficients in (3.1) are constant, and let $C = 0$. This gives the system

$$w_t = Aw_x + Bw_y . \tag{3.8}$$

If we insert a plane wave solution $v \exp(i\zeta x + i\omega y + i\xi t)$ into (3.8), where v is a vector, the result is $\xi v = (\zeta A + \omega B)v$. This is the condition (3.7) which was discussed earlier. The surfaces in graphs like Figure 3.1(a) therefore define the set of all possible frequencies for plane wave solutions to (3.8). It is apparent that the rapidly moving waves are associated with Ω, which is why we are interested in the behavior of (3.5) only for (ω,ξ) in Γ.

In graphs like Figure 3.1(a) there is a particular wave speed associated with each surface which defines ζ as a function of ω and ξ. This implies that it is possible to separate fast waves from slow waves by diagonalizing the symbol (3.5). It is also possible to detect the directions in which the fast waves are moving. The portion of the surface Ω for which the product $\xi\zeta$ is positive corresponds to waves which are leaving the region $x > 0$, and the portion for which $\xi\zeta < 0$ corresponds to waves which are entering the region. By properly defining the branches of ζ on the two sections of Γ, we can therefore separate the fast part of the solution into incoming and outgoing components. This justifies our decision to seek diagonal form for the symbol (3.5).

We need to say a little more about the directions in which the various waves propagate. A plane wave $\exp(i\zeta x + i\omega y + i\xi t)$ must propagate in the direction $\pm(\zeta,\omega)$. If the point (ω,ξ) lies on the vertical axis in Figure 3.1(b), then $\omega = 0$, and the wave moves in a direction normal to the boundary. If (ω,ξ) lies near the edge of Γ, then for a fast wave $|\zeta|$ is small compared to $|\omega|$, and the wave moves in a direction which is nearly tangential. This observation will be useful later when we seek explicit formulas for bringing about an approximate diagonalization of the symbol (3.5). The approximations we introduce will be valid asymptotically as $\frac{\omega}{\xi} \to 0$. This will lead to boundary conditions which work well for fast waves traveling in directions which have sizeable normal components, but they will not work well for waves moving in directions which are nearly tangential.

These tangential waves do not present any real problem, since they cannot influence the interior very rapidly. The approximation schemes are therefore worth using.

3.2 Outline of the Uncoupling Process

In this section we will describe the uncoupling process for systems in more than one space dimension. We will first outline some of the ideas using Fourier transforms in a formal way, and we will then use pseudo differential operators to make the process more rigorous.

We consider the system (3.1), $w_t = Aw_x + Bw_y + Cw$, on the domain $x > 0$. When we solve for w_x and apply Fourier transforms in y and t, the result is

$$\hat{w}_x(x,\omega,\xi) = (i\xi A^{-1} - i\omega A^{-1}B)\hat{w} - A^{-1}C\hat{w}. \tag{3.9}$$

According to the remarks of the preceding section, the leading symbol $i\xi A^{-1} - i\omega A^{-1}B$ is diagonalizable for (ω,ξ) in Γ, and it is only for (ω,ξ) in Γ that we can have rapidly moving waves. For the sake of neatness we will use a cutoff function to restrict attention to that set. Let φ be a C^∞ function of ω and ξ which is equal to zero outside Γ and which is equal to 1 on all of Γ except for a thin layer near the boundary. Equation (3.9) can be written

$$\hat{w}_x = (i\xi A^{-1} - i\omega\varphi A^{-1}B)\hat{w} - A^{-1}C\hat{w} - (1-\varphi)i\omega A^{-1}B\hat{w}. \tag{3.10}$$

The last term in (3.10) is an error term which is zero on almost all of Γ. It is nonzero only near the edge, and for fast waves this corresponds to nearly tangential incidence. The error term is therefore insignificant.

For any particular system it is necessary to find explicit formulas for similarity transformations which bring the leading symbol

(3.11) $$i\xi A^{-1} - i\omega\varphi A^{-1}B$$

of (3.10) to diagonal form, or at least to approximate diagonal form. We note that it would actually be enough to obtain a block diagonal form in which each block corresponds to slow, incoming fast, or outgoing fast components of the solution. This situation could occur with the Euler equations of gas dynamics, where there is a slow mode of multiplicity three. The uncoupling can be accomplished either by using a certain perturbation method or by explicitly computing the eigenvectors of (3.11).

To use the perturbation method we observe that (3.11) is equal to $i\xi$ times the matrix

(3.12) $$A^{-1} - (\tfrac{\omega}{\xi})\varphi A^{-1}B .$$

The matrix A has been assumed to be diagonal, so if $\frac{\omega}{\xi}$ is small, the matrix (3.12) is a perturbation of a diagonal matrix. We can therefore use the perturbation argument introduced in Chapter 2 to bring (3.12) closer to diagonal form, or at least to block diagonal form. We can apply the method once to reduce the coupling to order $(\frac{\omega}{\xi})^2$, twice to reduce it to order $(\frac{\omega}{\xi})^3$, and so forth. This is one of the approximation schemes mentioned in the preceding section which work well in directions having a sizeable normal component but do not work well near tangential incidence. We will present a general form of this perturbation method in the next section. There it will become apparent that in the case of multiple eigenvalues this

method cannot give diagonal form, but instead can give a satisfactory block diagonal form.

Another way to diagonalize (3.11) is to compute the eigenvectors explicitly. One way to do this is to work directly with the matrix (3.11). Another is to find the eigenvectors of the symbol (3.2), $\zeta A + \omega B$, and then use the ideas of Proposition 3.1 to translate these vectors into eigenvectors of (3.11). The latter approach would be preferable if (3.2) is easier to work with or if its eigenvectors are already known.

By calculating eigenvectors we will be able to obtain an exact diagonalization of (3.11) when (ω,ξ) is in Γ. This may appear to be an advantage over the perturbation method given earlier. However, the expressions for the eigenvectors can be complicated, and in order to obtain local boundary conditions it would usually be necessary to approximate these expressions with polynomials or rational functions. We would again use approximations which are valid asymptotically as $\frac{\omega}{\xi} \to 0$. Although it does not give exact results, the second approach allows greater flexibility in the choice of approximation methods. The earlier perturbation approach employs one fixed method of approximation, but here we have a choice of various Taylor approximations or rational Padé approximations. Engquist and Majda ([1], [2]) found Padé approximations particularly useful in their work on absorbing boundary conditions for scalar wave equations.

In the calculations for the shallow water equations which appear

in the next chapter, we will use the perturbation approach which was mentioned first. In this case the method gives satisfactory results. In general, however, one should keep in mind the greater flexibility allowed by the direct calculation of eigenvectors.

We now give an outline of the uncoupling process for systems in several space dimensions. Our intent at this point is to give a broad overview of the method and avoid details which could obscure the main ideas. We will go through the process in great detail in the next chapter when we derive boundary conditions for the shallow water equations. These calculations will be rather long and technical, so it will be worthwhile to first see a relatively short outline of the process.

We first solve for w_x in (3.1) to obtain the form (3.3),

$$w_x = A^{-1}w_t - A^{-1}Bw_y - A^{-1}Cw.$$

In order to simplify the notation we will change the meaning of A, B, and C and write the system as

$$w_x = Aw_t + Bw_y + Cw. \tag{3.13}$$

A, B, and C will have this meaning throughout the remainder of this paper. The matrix A is diagonal.

In order to prepare for the uncoupling, we will express (3.13) in terms of certain pseudo differential operators. As in Chapter 2 the solution w will be truncated in t so that these operators

can be applied to it, but this fact will be suppressed from the notation. Denote by H the operator with symbol

$$\sigma_H = i\xi A + i\omega\varphi B , \tag{3.14}$$

and let E_1 be the operator with symbol $i\omega(i-\varphi)B$. The system (3.13) then becomes

$$w_x = Hw + Cw + E_1 w . \tag{3.15}$$

As before, φ is a smooth function which is equal to 1 on almost all of Γ and is equal to zero outside Γ. The operator E_1 represents an error which is zero when $\varphi = 1$. In the case of variable coefficients the cone Ω, and therefore Γ, can vary with x, y, and t, so φ is in general a function of x, y, t, ω, and ξ.

We have mentioned that when we try to uncouple the leading symbol it will be useful to use perturbation arguments involving the quantity $\frac{\omega}{\xi}$. The operators which transform the system must therefore involve this quantity. A potential problem with this is that $\frac{\omega}{\xi}$ cannot be the symbol of a pseudo differential operator because of the singularity in the direction $\xi = 0$. However, we have avoided this difficulty by our use of the function φ. We will find that the ratio $\frac{\omega}{\xi}$ can appear only in the form $\frac{\omega}{\xi}\varphi$, and this is nonzero only in a conical neighborhood of the axis $\omega = 0$. The singularity is thereby eliminated.

When we uncouple the system (3.15) the first task is to take care of the leading order operator H. Let q be a matrix such that $q\sigma_H q^{-1}$ is approximately diagonal or approximately block diagonal,

and let Q be the pseudo differential operator whose symbol is q. The operator Q must have order zero, since its symbol q is homogeneous of order zero in its dependence on ω and ξ. If we apply Q to (3.15), the result is

$$(3.16) \qquad (Qw)_x = (QHQ^{-1})Qw + (QCQ^{-1} + Q_xQ^{-1})Qw + QE_1w .$$

Here Q^{-1} denotes a parametrix, or approximate inverse, of Q. This operator is defined by the property that $QQ^{-1} - I$ is a smoothing operator. It is not hard to show that such an operator exists and to obtain an asymptotic expansion for its symbol. An outline of the argument is given in Section 4.1. The leading order term in the expansion is q^{-1}, the inverse of the symbol of Q.

We need to examine the operator QHQ^{-1}. According to the composition law for pseudo differential operators, its leading symbol is the product of the leading symbols of Q, H, and Q^{-1}. This is the matrix $q\sigma_H q^{-1}$ which is known to be approximately uncoupled. There are also various lower order terms in the expansion of QHQ^{-1}. These are due partly to the effect of the composition law and partly to the lower order terms in the expansion for Q^{-1}. The composition law is stated in the Appendix.

Let G be the pseudo differential operator whose symbol is the diagonal part, or block diagonal part, of $q\sigma_H q^{-1}$, and let R be the operator whose symbol is the rest of $q\sigma_H q^{-1}$. G and R both have order one. Because of the approximate uncoupling in $q\sigma_H q^{-1}$,

the effect of R is small except near tangential incidence. If we let $w_0 = Qw$, (3.16) becomes

$$\frac{\partial w_0}{\partial x} = Gw_0 + Zw_0 + Rw_0 + E_2 w_0 . \tag{3.17}$$

Here Z is the pseudo differential operator associated with the zero-order terms appearing explicitly in (3.16) and with the terms of order zero or less which arise in the expansion of QHQ^{-1}. E_2 is equal to QE_1Q^{-1}, and its symbol is equal to a smoothing term when $\varphi = 1$. The system (3.17) is uncoupled near normal incidence, up to terms of order zero.

The coupling in the lower order term can be reduced by using the same technique that was used in Chapter 2 for systems in one space dimension. That is, we can apply to (3.17) an operator of the form $I + K$, where I is the identity operator and K is an operator of order -1 which is to be determined. When we apply the operator $I + K$, the result is

$$\frac{\partial}{\partial x}\left[(I+K)w_0\right] = (I+K)G(I+K)^{-1}w_1 + (I+K)(Z+R+E_2)(I+K)^{-1}w_1 + K_x w_0 , \tag{3.18}$$

where $w_1 = (I+K)w_0 = (I+K)Qw$. The parametrix $(I+K)^{-1}$ has the asymptotic expansion $I - K + K^2 - \ldots$. This follows easily from the fact that the order of K is negative. The system (3.18) can therefore be written

$$\frac{\partial w_1}{\partial x} = Gw_1 + (KG - GK + Z)w_1$$

$$+ \text{(terms of order -1 or less)} \tag{3.19}$$

$$+ (K+K)(R+E_2)(I+K)^{-1}w_1 .$$

The zero-order coupling in (3.19) is caused by the operator $KG - GK + Z$. Its leading symbol is

$$\sigma_K\sigma_G - \sigma_G\sigma_K + z_0 , \tag{3.20}$$

where σ_K and σ_G are the symbols of K and G, and z_0 is the leading symbol of Z. In order to eliminate the coupling of order zero, we will need to determine σ_K so that the symbol (3.20) is diagonal (or block diagonal). In Chapter 2 we did this calculation for a special case, and in the next section we will give a more general treatment as part of a general perturbation lemma. There we will find that it is possible to find a suitable σ_K provided that the diagonal blocks of σ_G have disjoint spectra. This condition can be satisfied here since we are trying to separate slow, incoming fast, and outgoing fast components of the solution.

The technique given here can be used to uncouple the system further. To reduce the coupling from order $-n+1$ to order $-n$, we would use an operator of the form $I + K_n$, where K_n has order $-n$. After m uncouplings the dependent variable would be

$$w_m = (I+K_m) \cdot \ldots \cdot (I+K_1)Qw ,$$

where K_1 is the operator K discussed above.

Boundary conditions for the system can be generated and then analyzed using ideas which are similar to those used in the case of one space dimension. To do the analysis we would localize the solution to a neighborhood of a bicharacteristic and then find energy estimates involving Sobolev norms. These estimates would give information about the behavior of the solution at high frequencies. The fact that the solution can be localized to a bicharacteristic means that we can study the effects of the boundary conditions for various angles of incidence to the boundary. This gives meaning to the use of the approximations about normal incidence which were mentioned earlier.

3.3 A Perturbation Lemma

In this section we present a method for reducing the coupling found in matrices which are perturbations of block diagonal matrices. This method can be used to partially uncouple the leading symbol in the system (3.13), and it is essentially the method which has already been used to reduce the coupling caused by lower order terms. We present it as a separate lemma for the sake of clarity and generality. Various versions of this method have been used in [4], [8], and [10] for 2×2 block matrices.

Proposition 3.2. Let A and B be square matrices of equal dimension. Suppose that A is block diagonal, and let $A_1, \ldots, A_n$ denote the blocks on the diagonal. If no two of the A_j have any eigenvalues in common, then for small ε the sum $A + \varepsilon B$ can be uncoupled to order ε^2. More precisely, there exists a matrix M such that for ε sufficiently small,

$$(I+\varepsilon M)(A+\varepsilon B)(I+\varepsilon M)^{-1} = A + \varepsilon \cdot (\text{block diagonal matrix}) + \mathcal{O}(\varepsilon^2) .$$

A method for constructing M will be given in the proof.

Proof. For small ε the inverse $(I+\varepsilon M)^{-1}$ exists and is equal to $I - \varepsilon M + \varepsilon^2 M^2 - \ldots$. We can therefore write

$$\begin{aligned}(I+\varepsilon M)(A+\varepsilon B)(I+\varepsilon M)^{-1} &= (I+\varepsilon M)(A+\varepsilon B)(I-\varepsilon M + \mathcal{O}(\varepsilon^2)) \\ &= A + \varepsilon(MA-AM+B) + \mathcal{O}(\varepsilon^2) .\end{aligned}$$

Our goal is to choose M so that

$$MA - AM + B \tag{3.21}$$

is block diagonal. For the sake of notation we will partition M and B into block structures which match the block structure of A. M_{ij} and B_{ij} will denote the blocks in the (i,j) position. They are not necessarily square, since we are not assuming that A_i and A_j have the same dimensions. The (i,j) block in (3.21) can then be written as $M_{ij}A_j - A_iM_{ij} + B_{ij}$. For $i \neq j$ we want this to be equal to zero. We are therefore faced with the problem of solving the equation

$$M_{ij}A_j - A_iM_{ij} = -B_{ij} \tag{3.22}$$

for M_{ij}. Once we have done this, the proof is complete. There are no conditions imposed on the diagonal blocks M_{ii}, so these may be chosen arbitrarily.

If A_i and A_j are both 1×1 matrices, i.e., scalars, then we obviously need to have $A_i \neq A_j$ in order to be able to solve (3.22) for arbitrary B_{ij}. In the general case the system (3.22) is solvable if and only if A_i and A_j have disjoint spectra. Proofs of this fact can be found in several different references. We give one here for the sake of completeness.

In order to simplify the notation we will write (3.22) in the form $XS - TX = Y$, where S, T, and Y are given and X is to be determined. We are assuming that S and T are square matrices which do not have any eigenvalues in common. There is no need to assume that

they have the same dimension. We will denote the columns of X and Y by x_j and y_i, and we will denote the entries of S by s_{ij}.

We can assume that S is upper triangular, since otherwise we can use a similarity transformation to reduce the problem to that case. We will solve for the columns of X, starting from the left. We first have $s_{11}x_1 - Tx_1 = y_1$. The matrix $s_{11}I - T$ is nonsingular since s_{11} is an eigenvalue of S and therefore not an eigenvalue of T. The column x_1 is therefore determined uniquely. We next have $s_{22}x_2 - Tx_2 = y_2 - s_{12}x_1$. This system has a unique solution x_2 since s_{22} is not an eigenvalue of T. We can continue in this manner to solve for all of X. We note that the condition on the eigenvalues of S and T is necessary as well as sufficient, since it is equivalent to the statement that $s_{ii}I - T$ is nonsingular for all i. This completes the proof of the lemma, and therefore the proof of the main proposition.

CHAPTER 4

AN EXAMPLE IN TWO SPACE DIMENSIONS

In this chapter we will use the methods of Chapter 3 to derive boundary conditions for the linearized shallow water equations. The calculations will follow the outline given in Section 3.2. When we uncouple the leading symbol of the system we will use the perturbation method given in Section 3.3. This process and the one used to reduce the lower order coupling will each be applied one time. Certain portions of the calculations are specific to the shallow water equations, but other portions are more generally applicable. For much of the chapter the spatial domain we consider will be the half-space $x > 0$, but later we will discuss the effect of rotation of coordinates on the form of the boundary conditions. In the last section we will present the results of some numerical tests of these conditions.

4.1 Uncoupling the System

The linearized shallow water equations can be written in the form

$$(4.1)\quad \frac{\partial}{\partial t}\begin{pmatrix} u \\ v \\ p \end{pmatrix} = \begin{pmatrix} \alpha & & -c \\ & \alpha & \\ -c & & \alpha \end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} u \\ v \\ p \end{pmatrix} + \begin{pmatrix} \beta & & \\ & \beta & -c \\ & -c & \beta \end{pmatrix}\frac{\partial}{\partial y}\begin{pmatrix} u \\ v \\ p \end{pmatrix}$$

$$+ (\gamma_{ij})\begin{pmatrix} u \\ v \\ p \end{pmatrix}$$

In this notation, $(-\alpha,-\beta)$ is the velocity of the flow about which the system has been linearized, and c is a speed associated with the propagation of gravity waves. We will assume $|\alpha|,|\beta| << c$ and $\alpha \neq 0$. The dependent variables are given by $u = cu'$, $v = cv'$, and $p = \phi'$, where u' and v' are the perturbations in the components of velocity and ϕ' is the perturbation in the geopotential. The coefficients γ_{ij} in the undifferentiated term are due partly to Coriolis effects and partly to the process of linearization. For the time being we will consider the system on the domain $x > 0$.

The first step is to diagonalize the coefficient matrix of the normal derivative $\frac{\partial}{\partial x}$ in (4.1). To do this we use the matrix

$$\frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 & 1 \\ \sqrt{2} & 0 & 0 \\ 0 & 1 & -1 \end{pmatrix} . \tag{4.2}$$

The columns of this matrix are normalized eigenvectors of the coefficient matrix in question. When we multiply (4.1) on the left by the inverse of (4.2) and make the appropriate change of dependent variable, the result is the system

$$w_t = \begin{pmatrix} \alpha & & \\ & \alpha-c & \\ & & \alpha+c \end{pmatrix} w_x + Dw_y + Ew , \tag{4.3}$$

where

$$w = \begin{pmatrix} v \\ \frac{1}{\sqrt{2}}(u+p) \\ \frac{1}{\sqrt{2}}(u-p) \end{pmatrix} . \tag{4.4}$$

D and E can be written explicitly, but we will wait until we solve for w_x.

We can use (4.3) to make a preliminary identification of the slow, incoming fast, and outgoing fast parts of the solution. Since $|\alpha| << c$, we can say roughly that the first component v is a slow component and that $u+p$ and $u-p$ are incoming and outgoing fast components, respectively. In order to suppress the incoming fast part of the solution we could therefore require $u+p=0$ at the boundary $x=0$. If $\alpha < 0$, then we would also prescribe a value for v in order to have a well-posed problem. The trouble with this approach is that it ignores the effect of the terms Dw_y and Ew. Our identification of the various parts of the solution is therefore not very accurate. The purpose of the uncoupling process is to produce a more accurate identification and thereby enable us to find boundary conditions which are more effective at suppressing the incoming fast part.

We need to solve for w_x in (4.3). When we do this the result is

$$w_x = Aw_t + Bw_y + Cw,$$

where

$$A = \begin{pmatrix} \frac{1}{\alpha} & & \\ & \frac{1}{\alpha-c} & \\ & & \frac{1}{\alpha+c} \end{pmatrix}$$

(4.6)
$$B = \begin{pmatrix} \frac{-\beta}{\alpha} & \frac{c}{\alpha\sqrt{2}} & \frac{-c}{\alpha\sqrt{2}} \\ \frac{c}{\sqrt{2}\ (\alpha-c)} & \frac{-\beta}{\alpha-c} & 0 \\ \frac{-c}{\sqrt{2}\ (\alpha+c)} & 0 & \frac{-\beta}{\alpha+c} \end{pmatrix}$$

$$C = -\begin{pmatrix} \frac{\gamma_{22}}{\alpha} & \frac{\gamma_{21}+\gamma_{23}}{\alpha\sqrt{2}} & \frac{\gamma_{21}-\gamma_{23}}{\alpha\sqrt{2}} \\ \frac{\gamma_{12}+\gamma_{32}}{\sqrt{2}\ (\alpha-c)} & \frac{\gamma_{11}+\gamma_{13}+\gamma_{31}+\gamma_{33}}{2(\alpha-c)} & \frac{\gamma_{11}-\gamma_{13}+\gamma_{31}-\gamma_{33}}{2(\alpha-c)} \\ \frac{\gamma_{12}-\gamma_{32}}{\sqrt{2}\ (\alpha+c)} & \frac{\gamma_{11}+\gamma_{13}-\gamma_{31}-\gamma_{33}}{2(\alpha+c)} & \frac{\gamma_{11}-\gamma_{13}-\gamma_{31}+\gamma_{33}}{2(\alpha+c)} \end{pmatrix}.$$

As in Section 3.2 we write the system (4.5) in the form

(4.7)
$$w_x = Hw + Cw + E_1 w,$$

where H is the operator whose symbol is given by

(4.8)
$$\sigma_H = i\xi A + i\omega\varphi B,$$

and E_1 is the operator with symbol $i\omega(1-\varphi)B$. In order to uncouple the leading order part of (4.7) we need to find a symbol q such that

$q\sigma_H q^{-1}$ is closer to diagonal form than σ_H, at least for small $\frac{\partial}{\xi}$. To do this we will use the ideas of sections 3.2 and 3.3 to find a matrix M such that

$$(I + \tfrac{\partial}{\xi}\varphi M)(A + \tfrac{\partial}{\xi}\varphi B)(I + \tfrac{\partial}{\xi}\varphi M)^{-1}$$

$$= \text{diagonal matrix} + (\mathcal{O}(\tfrac{\partial}{\xi}\varphi)^2) \quad . \tag{4.9}$$

We will then let $q = I + \frac{\partial}{\xi}\varphi M$.

Satisfying the condition (4.9) amounts to solving the equation $MA - AM + B = 0$ for M. A calculation shows that M can be taken to be

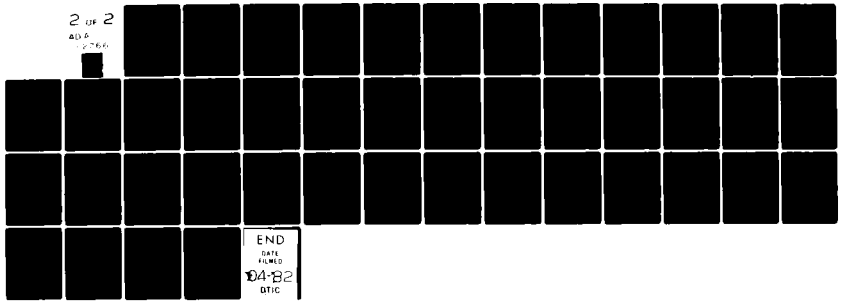
$$M = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & -(\alpha-c) & -(\alpha+c) \\ \alpha & 0 & 0 \\ \alpha & 0 & 0 \end{pmatrix}$$

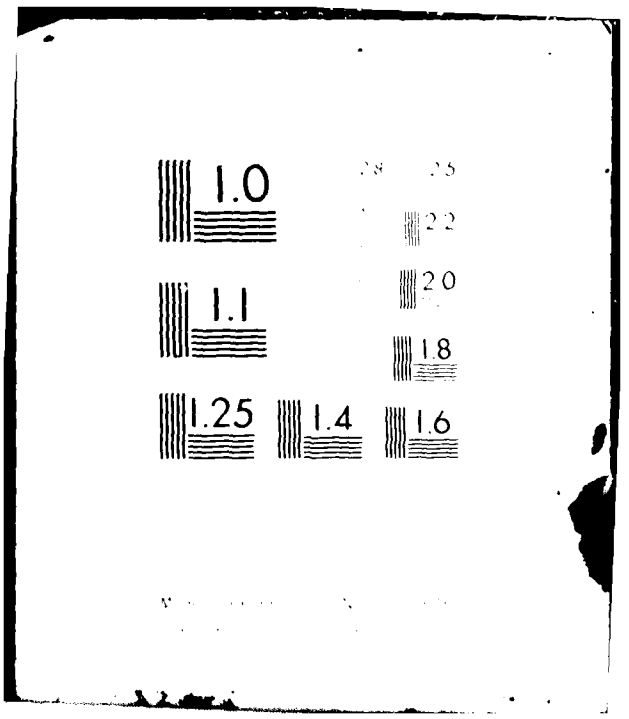
and that $(I + \varepsilon M)(A + \varepsilon B)(I + \varepsilon M)^{-1}$ is then equal to

$$\begin{pmatrix} \frac{1}{\alpha} & & \\ & \frac{1}{\alpha-c} & \\ & & \frac{1}{\alpha+c} \end{pmatrix} + \varepsilon \begin{pmatrix} -\frac{\beta}{\alpha} & & \\ & -\frac{\beta}{\alpha-c} & \\ & & -\frac{\beta}{\alpha+c} \end{pmatrix} + \mathcal{O}(\varepsilon^2) \tag{4.11}$$

The off-diagonal elements in (4.10) are determined uniquely by the condition (4.9), but the diagonal elements may be chosen arbitrarily. For convenience we have set these equal to zero.

AD-A112 766 STANFORD UNIV CA DEPT OF COMPUTER SCIENCE F/G 20/4
BOUNDARY CONDITIONS FOR HYPERBOLIC SYSTEMS OF PARTIAL DIFFERENT--ETC(U)
AUG 81 R L HIGDON N00014-75-C-1132
UNCLASSIFIED STAN-CS-81-890 NL




1.0
1.1
1.25
1.4
1.6
2.2
2.0
1.8

We now define the symbol q by

$$q = I + \frac{\omega}{\xi}\varphi M, \tag{4.12}$$

where M is given in (4.10), and we let Q be the pseudo differential operator with symbol q. When the operator Q is applied to the system (4.7), the result is

$$(Qw)_x = (QHQ^{-1})Qw + (QCQ^{-1} + Q_xQ^{-1})Qw + QE_1w. \tag{4.13}$$

Here Q_x denotes the operator with symbol q_x.

We saw in Section 3.2 that the leading symbol of QHQ^{-1} is $q\sigma_H q^{-1}$. According to (4.8) and (4.12) this is given by

$$q\sigma_H q^{-1} = (I + \tfrac{\omega}{\xi}\varphi M)(i\xi A + i\omega\varphi B)(I + \tfrac{\omega}{\xi}\varphi M)^{-1}.$$

If we factor out $i\xi$ and identify $\frac{\omega}{\xi}\varphi$ with ε in (4.11), we can conclude that $q\sigma_H q^{-1}$ is equal to $\sigma_G + (\frac{\omega^2}{\xi}\varphi^2)$, where

$$\sigma_G = i\xi \begin{pmatrix} \frac{1}{\alpha} & & \\ & \frac{1}{\alpha-c} & \\ & & \frac{1}{\alpha+c} \end{pmatrix} + i\omega\varphi \begin{pmatrix} \frac{-\beta}{\alpha} & & \\ & \frac{-\beta}{\alpha-c} & \\ & & \frac{-\beta}{\alpha+c} \end{pmatrix}. \tag{4.14}$$

We will let G denote the operator whose symbol is σ_G. The system (4.13) can then be written in the form

$$\frac{\partial w_0}{\partial x} = Gw_0 + Zw_0 + \mathcal{O}\left(\frac{\omega^2}{\xi}\varphi^2\right)w + \mathcal{O}(\omega(1-\varphi))w, \tag{4.15}$$

where $w_0 = Qw$.

The operator Z is associated with the zero-order terms appearing explicitly in (4.13) and with the terms of order zero or less which arise in the expansion of the symbol of QHQ^{-1}. In a moment we will discuss this further. The term $\mathcal{O}\left(\frac{\omega^2}{\xi}\varphi^2\right)w$ denotes the effect of an operator whose symbol is dominated by $\frac{\omega^2}{\xi}\varphi^2$ and which is a result of the error in uncoupling the leading order part of the system (4.7). The term $\mathcal{O}(\omega(1-\varphi))w$ represents the effect of the operator E_1 which appears in (4.13). Its symbol is equal to zero on almost all of the set Γ, which is the only part of the (ω,ξ) space in which fast waves can be found. In Γ it is nonzero only near the edge, which for fast waves corresponds to nearly tangential incidence. This term is therefore of no consequence.

The system is now partially uncoupled near normal incidence, since the symbol of G is diagonal. We next need to reduce the coupling caused by the zero-order operator Z, which is given by

$$Z = QCQ^{-1} + Q_xQ^{-1} + (\text{terms of order zero or less arising from the expansion of } QHQ^{-1}). \tag{4.16}$$

The coupling can be reduced by the method presented in Chapters 2 and 3, but we will first have to identify the leading symbol of Z.

We first consider the first two terms in (4.16). The leading symbol of QCQ^{-1} is

$$qCq^{-1} = (I + \tfrac{3}{\xi}\varphi M)C(I + \tfrac{3}{\xi}\varphi M)^{-1}$$

$$= C + \mathcal{O}(\tfrac{3}{\xi}\varphi) ,$$

and the leading symbol of Q_xQ^{-1} is

$$q_xq^{-1} = (\tfrac{3}{\xi}\varphi M_x)(I + \tfrac{3}{\xi}\varphi M)^{-1}$$

$$= \mathcal{O}(\tfrac{3}{\xi}\varphi) .$$

The expression for q is taken from (4.12). We will regard terms of order $\frac{3}{\xi}\varphi$ as error terms since $\frac{3}{\xi}\varphi$ is no larger than the term $\frac{3^2}{\xi}\varphi^2$ which has already appeared in (4.15). The first two terms in (4.16) are then given by

$$QCQ^{-1} + Q_xQ^{-1} = C + \text{order}(-1) + \mathcal{O}(\tfrac{3}{\xi}\varphi) \tag{4.17}$$

In order to consider the expansion of QHQ^{-1} we must first find the symbol of the parametrix Q^{-1}. We start by assuming that the symbol has an asymptotic expansion of the form r_0 , where r_k is homogeneous of order $-k$ in ω and ξ. The choice of orders will turn out to be appropriate, since Q has order zero.

We will then solve for the r_k one by one.

Let R_k be the operator with symbol r_k. According to the composition law for pseudo differential operators, the symbol of QR_0 is

$$qr_0 + \frac{1}{i}\frac{\partial q}{\partial \xi}\frac{\partial r_0}{\partial t} + \frac{1}{i}\frac{\partial q}{\partial \omega}\frac{\partial r_0}{\partial y} + \text{order } (-2) .$$

The composition law is stated in the Appendix. If we choose $r_0 = q^{-1}$, then the leading term is I. We note that r_0 really is of order zero and that the error term really is of order -2. We now have

$$QR_0 = I + \text{order } (-1) . \tag{4.18}$$

Now choose r_1 so that

$$qr_1 = -\frac{1}{i}\left(q_\xi \frac{\partial r_0}{\partial t} + q_\omega \frac{\partial r_0}{\partial y} \right). \tag{4.19}$$

The leading symbol of QR_1 therefore cancels the leading symbol of the error in (4.18). This gives

$$Q(R_0 + R_1) = I + \text{order } (-2) .$$

This process can be continued indefinitely to find any r_k. At each step we would need to know the leading symbol of the error in the equation

$$Q(R_0 + \ldots + R_{k-1}) = I + \text{order } (-k) .$$

This can be calculated from the expansions of the symbols of QR_j for $j < k$.

In the present problem we really only need the terms r_0 and r_1. Q has order zero and H has order one, so the terms $r_2, r_3, \ldots$ must contribute terms of negative order in the expansion of QHQ^{-1}. These lower order terms are of no interest to us here.

The leading symbol r_0 of Q^{-1} is given by

$$r_0 = q^{-1} = (I + \frac{\omega}{\xi}\varphi M)^{-1} = I + \ell(\frac{\omega}{\xi}\varphi). \tag{4.20}$$

We will see later that in this case it is not necessary to keep any more terms in this expansion of r_0. The second term r_1 in Q^{-1} is defined in (4.19) to be

$$r_1 = iq^{-1}(q_\xi \frac{\partial r_0}{\partial t} + q_\omega \frac{\partial r_0}{\partial y}).$$

Equations (4.12) and (4.20) can be used to obtain

$$\begin{aligned} r_1 &= i(I + \tfrac{\omega}{\xi}\varphi M)^{-1}[(-\tfrac{\omega}{\xi^2}\varphi M)\,\ell'(\tfrac{\omega}{\xi}\varphi) + (\tfrac{1}{\xi}\varphi M)\ell(\tfrac{\omega}{\xi}\varphi)] \\ &= \ell'(\tfrac{\omega}{\xi^2}\varphi^2) \end{aligned} \tag{4.21}$$

In (4.21) we have omitted derivatives of φ since these are nonzero only near the edge of Γ and cannot be of any consequence. From the above work we can conclude that the symbol of Q^{-1} has the expansion

$$r_0 + r_1 + \text{order}(-2)$$

$$= q^{-1} + r_1 + \text{order}(-2)$$

$$= I + \mathcal{O}(\frac{\omega}{\xi}\varphi) + \mathcal{O}(\frac{\omega}{\xi^2}\varphi^2) + \text{order}(-2) .$$

We are now ready to calculate the expansion for QHQ^{-1}. We will begin by finding the symbol of HQ^{-1}. According to the composition law, this is

$$\sigma_{HQ^{-1}} = \sigma_H(r_0 + r_1 + \text{order}(-2)) + \frac{1}{i}\frac{\partial\sigma_H}{\partial\xi}\frac{\partial}{\partial t}(r_0 + \text{order}(-1))$$

$$+ \frac{1}{i}\frac{\partial\sigma_H}{\partial\omega}\frac{\partial}{\partial y}(r_0 + \text{order}(-1)) + \text{order}(-1) .$$

Since $r_0 = q^{-1}$ and $\sigma_H = i\xi A + i\omega\varphi B$, this can be written as

$$\sigma_{HQ^{-1}} = \sigma_H q^{-1} + (i\xi A + i\omega\varphi B)r_1 + A\frac{\partial r_0}{\partial t} + \varphi B\frac{\partial r_0}{\partial y} + \text{order}(-1) .$$

We have again omitted derivatives of φ. (4.20) and (4.21) now imply

$$\sigma_{HQ^{-1}} = \sigma_H q^{-1} + (i\xi A + i\omega\varphi B)\mathcal{O}(\frac{\omega}{\xi^2}\varphi^2)$$

$$+ A\mathcal{O}(\frac{\omega}{\xi}\varphi) + \varphi B\mathcal{O}(\frac{\omega}{\xi}\varphi) + \text{order}(-1) ,$$

or

$$\sigma_{HQ^{-1}} = \sigma_H q^{-1} + \mathcal{O}(\frac{\omega}{\xi}\varphi) + \text{order}(-1) . \tag{4.22}$$

The symbol of QHQ^{-1} is given by

$$\sigma_{QHQ^{-1}} = q\sigma_{HQ^{-1}} + \frac{1}{i}\frac{\partial q}{\partial \xi}\frac{\partial}{\partial t}\sigma_{HQ^{-1}} + \frac{1}{i}\frac{\partial q}{\partial \omega}\frac{\partial}{\partial y}\sigma_{HQ^{-1}} + \text{order}(-1) .$$

We now use (4.22) and the fact that q is equal to $I + \frac{\omega}{\xi}\varphi M$.

$$\sigma_{QHQ^{-1}} = q[\sigma_H q^{-1} + \ell(\tfrac{\omega}{\xi}\varphi) + \text{order}(-1)]$$

$$+ \frac{1}{i}\left(-\frac{\omega}{\xi^2}\varphi M\right)\frac{\partial}{\partial t}\sigma_{HQ^{-1}} \tag{4.23}$$

$$+ \frac{1}{i}\left(\frac{1}{\xi}\varphi M\right)\frac{\partial}{\partial y}\sigma_{HQ^{-1}} + \text{order}(-1) .$$

A short calculation shows that

$$\frac{\partial}{\partial t}\sigma_{HQ^{-1}} = (i\xi A_t + i\omega\varphi B_t)(I - \tfrac{\omega}{\xi}\varphi M)$$

$$+ (i\xi A + i\omega\varphi B)(-\tfrac{\omega}{\xi}\varphi M_t)$$

$$+ \ell(\tfrac{\omega^2}{\xi}\varphi^2) + \ell(\tfrac{\omega}{\xi}\varphi) + \text{order}(-1).$$

The same relation holds when we replace t with y. The second term in (4.23) is therefore $\ell(\frac{\omega}{\xi}\varphi) + \text{order}(-1)$, and the third is $\varphi M A_y + \ell(\frac{\omega}{\xi}\varphi) + \text{order}(-1)$. Equation (4.23) can then be simplified to

$$(4.24) \qquad \sigma_{QHQ^{-1}} = q\sigma_H q^{-1} + \varphi MA_y + \mathcal{O}(\tfrac{\omega}{\xi}\varphi) + \text{order } (-1) .$$

The first term is the term of order one which we have already used, and the second is the term of order zero which we have been seeking.

The point of all of this work was to find the symbol, at least to leading order, of the operator Z which represents the zero-order coupling in the system (4.15). According to (4.16), this operator is given by

$$Z = QCQ^{-1} + Q_x Q^{-1} + (\textit{terms from } QHQ^{-1} \textit{ of order zero or less}) .$$

Our results in (4.17) and (4.24) imply that its symbol is

$$(4.25) \qquad C + \varphi MA_y + \mathcal{O}(\tfrac{\omega}{\xi}) + \text{order } (-1) .$$

The system (4.15) can therefore be written in the form

$$(4.26) \qquad \frac{\partial w_0}{\partial x} = Gw_0 + Z_0 w_0 + \mathcal{O}(\tfrac{\omega^2}{\xi}\varphi)w + \mathcal{O}(\tfrac{\omega}{\xi}\varphi)w + \mathcal{O}(w(1-\varphi))w + (\text{order } -1)w ,$$

where Z_0 is the operator whose symbol is $\varphi(C + MA_y)$. We have split the matrix C into the sum $\varphi C + (1-\varphi)C$ in order to give a neater form to certain formulas which will appear later.

We are finally ready to uncouple the term of order zero. To do this we will use the method given near the end of Section 3.2. That is, we will apply an operator of the form $I + K$ to (4.26), where K

has order -1, and then make the corresponding change of dependent variable. We will choose K so that the zero-order operator in the transformed system has a diagonal leading symbol. According to the work in Section 3.2, this operator is $KG - GK + Z_0$. We therefore need to choose K so that

$$\sigma_K\sigma_G - \sigma_G\sigma_K + \varphi(C + MA_y) = \text{diagonal matrix}. \tag{4.27}$$

Here σ_K and σ_G are the symbols of K and G, respectively, and the third term is the symbol of Z_0. σ_G is given in (4.15), A and C are given in (4.6), and M is given in (4.10).

We can solve (4.27) using the method of Proposition 3.2. After a certain amount of labor we obtain

$$\sigma_K = \frac{\varphi}{i\xi(1 - \frac{\omega}{\xi}\varphi\beta)c} \begin{pmatrix} 0 & k_{12} & k_{13} \\ k_{21} & 0 & k_{23} \\ k_{31} & k_{32} & 0 \end{pmatrix}, \tag{4.28}$$

where

$$k_{12} = \frac{1}{\sqrt{2}}\left[(\gamma_{21} + \gamma_{23})(\alpha - c) - \alpha(\alpha_y - c_y)\right]$$

$$k_{13} = \frac{1}{\sqrt{2}}\left[(-\gamma_{21} + \gamma_{23})(\alpha + c) + \alpha(\alpha_y + c_y)\right]$$

$$k_{21} = \frac{1}{\sqrt{2}}\left[-\alpha(\gamma_{12} + \gamma_{32}) - \alpha_y(\alpha - c)\right]$$

$$k_{31} = \frac{1}{\sqrt{2}}\left[\alpha(\gamma_{12} - \gamma_{32}) + \alpha_y(\alpha + c)\right]$$

$$k_{23} = \frac{1}{4}(-\gamma_{11} + \gamma_{13} - \gamma_{31} + \gamma_{33})(\alpha + c)$$

$$k_{32} = \frac{1}{4}(\gamma_{11} + \gamma_{13} - \gamma_{31} - \gamma_{33})(\alpha - c).$$

Equation (4.27) does not impose any conditions on the diagonal elemnets of K. For convenience we have set these equal to zero.

The operator $I + K$ transforms the system (4.26) into the form

$$(4.29)\quad \frac{\partial w_1}{\partial x} = Gw_1 + (\text{diagonal operator of order zero})w_1 + (\text{order}(-2))w + \mathcal{O}\left(\frac{\omega^2}{\xi}\varphi\right)w + \mathcal{O}\left(\frac{\omega}{\xi}\varphi\right)w + \mathcal{O}(\omega(1-\varphi))w\,,$$

where $w_1 = (I+K)w_0 = (I+K)Qw$. The symbol of K is given in (4.28), the symbol of Q is given in (4.12), and the components of w are given in (4.4). This represents all of the uncoupling which we will do for this system.

4.2 Boundary Conditions

It is now time to use the results of the uncoupling process to derive boundary conditions for the system (4.1). It is necessary to identify the incoming fast component for the partially uncoupled system (4.29) and then find conditions which suppress this component at the boundary $x = 0$.

The symbol of the operator G which appears in (4.29) is given in (4.14) and is equal to

$$\sigma_G = i\xi \begin{pmatrix} \frac{1}{\alpha} & & \\ & \frac{1}{\alpha - c} & \\ & & \frac{1}{\alpha + c} \end{pmatrix} + i\omega\varphi \begin{pmatrix} \frac{-\xi}{\alpha} & & \\ & \frac{-\xi}{\alpha - c} & \\ & & \frac{-\xi}{\alpha + c} \end{pmatrix}.$$

Since $|\alpha \pm c| >> |\alpha|$, the second and third components of w_1 in (4.29) are the rapidly moving portions of the solution. The second is the incoming component, since $\alpha - c < 0$. We need to use the identity $w_1 = (I+K)Qw$ to find an explicit formula for this component, and then we need to use this formula to find suitable boundary conditions for (4.1).

The dependent variable w_1 is given by

$$\begin{aligned} w_1 &= (I+K)Qw \\ &= \text{"}(I + \sigma_k)\text{"} \circ \text{"}(I + \tfrac{\omega}{\xi}\varphi M)\text{"}w . \end{aligned} \tag{4.30}$$

Here we have used the expression (4.12) for the symbol of Q. The quote marks denote pseudo differential operators having the stated

symbols, and the small circle denotes composition of operators. (4.30) can be written

$$w_1 = "(\varphi + \sigma_K)" \circ "(\varphi + \tfrac{\omega}{\xi}\varphi M)"w + \mathcal{O}(1-\varphi) . \tag{4.31}$$

In order to produce cleaner formulas later on we have used the cut-off function φ to restrict the solution to the set Γ in the (ω,ξ) space in which the fast waves can be found.

According to the composition law, the symbol of the composition in (4.31) is

$$\begin{aligned}&(\varphi + \sigma_K)(\varphi + \tfrac{\omega}{\xi}\varphi M) + \frac{1}{i}\frac{\partial}{\partial\xi}(\varphi + \sigma_K)\frac{\partial}{\partial t}(\varphi + \tfrac{\omega}{\xi}\varphi M)\\ &+ \frac{1}{i}\frac{\partial}{\partial\omega}(\varphi + \sigma_K)\frac{\partial}{\partial y}(\varphi + \tfrac{\omega}{\xi}\varphi M) + \text{order } (-2) .\end{aligned} \tag{4.32}$$

The derivatives of $\varphi + \sigma_K$ with respect to ω and ξ are of order -2, since σ_K has order -1 and we are ignoring derivatives of φ. (4.32) is therefore

$$\begin{aligned}&(\varphi + \sigma_K)(\varphi + \tfrac{\omega}{\xi}\varphi M) + \text{order } (-2)\\ &= \varphi^2 + \varphi\sigma_K + \tfrac{\omega}{\xi}\varphi^2 M + \tfrac{\omega}{\xi}\varphi\sigma_K M + \text{order } (-2)\\ &= \varphi^2 + \varphi\sigma_K + \tfrac{\omega}{\xi}\varphi^2 M + \mathcal{O}\left(\tfrac{\omega}{\xi^2}\varphi^2\right) + \text{order } (-2) .\end{aligned} \tag{4.33}$$

To obtain the last line we used (4.28) to conclude $\sigma_K = \mathcal{O}(\frac{1}{\xi}\varphi)$. (4.31) and (4.33) now imply

$$(4.34)\quad w_1 = \text{"}(\varphi^2 + \varphi\sigma_K + \tfrac{\omega}{\xi}\varphi^2 M)\text{"}w + \mathcal{O}(\tfrac{\omega}{\xi^2}\varphi^2) + \mathcal{O}(1-\varphi) + \text{order}(-2)\ .$$

The error terms in (4.34) can be ignored in the system (4.29), since their only contributions in that equation are error terms having the same order as terms which are already there. In particular, $G \circ \mathcal{O}(\frac{\omega}{\xi^2}\varphi^2) = (\mathcal{O}(\xi) + \mathcal{O}(\omega)) \circ \mathcal{O}(\frac{\omega}{\xi^2}\varphi^2) = \mathcal{O}(\frac{\omega}{\xi}\varphi^2) + \mathcal{O}(\frac{\omega^2}{\xi^2}\varphi^2)$. The system (4.29) can therefore be written

$$(4.35)\quad \frac{\partial z}{\partial x} = Gz + (\text{diagonal of order zero})z + (\text{order}(-1))w + (\tfrac{\omega^2}{\xi}\varphi)w + (\tfrac{\omega}{\xi}\varphi)w + (\omega(1-\varphi))w\ ,$$

where

$$(4.36)\qquad z = \text{"}(\varphi^2 + \varphi\sigma_K + \tfrac{\omega}{\xi}\varphi^2 M)\text{"}w$$

σ_K is given in (4.28) and M is given in (4.10).

The second component of z is the incoming fast component which we need to suppress. According to (4.36) and the expressions for σ_K and M, this component is

$$z_2(x,y,t) = \int e^{i(\omega y+\xi t)}\varphi^2 \left\{ \left[(\tfrac{\omega}{\xi})\frac{\alpha}{\sqrt{2}} - (\frac{1}{c\sqrt{2}})\frac{\alpha(\gamma_{12}+\gamma_{32}) + \lambda_\nu(\alpha-c)}{i\xi(1 - \frac{\omega}{\xi}\varphi 3)} \right] \hat{w}(1) \right.$$
$$(4.37)\qquad \left. + \hat{w}(2) - (\frac{1}{4c})\frac{(\alpha+c)(\gamma_{11}-\gamma_{13}+\gamma_{31}-\gamma_{33})}{i\xi(1 - \frac{\omega}{\xi}\varphi 3)}\ \hat{w}(3) \right\} d\omega d\xi\ .$$

Here $w^{(1)}$, $w^{(2)}$, and $w^{(3)}$ denote the components of w. These are given explicitly in (4.4). We note that (4.37) is a perturbation of $w^{(2)}$.

This expression can be simplified somewhat. The factor $(1 - \frac{\omega}{\xi}\varphi\beta)^{-1}$ can be approximated by $1 + \mathcal{O}(\frac{\omega}{\xi}\varphi)$. When this is multiplied by $(i\xi)^{-1}$ the result is $(i\xi)^{-1} + \mathcal{O}(\frac{\omega}{\xi^2}\varphi)$. For reasons stated earlier, terms of order $\frac{\omega}{\xi^2}\varphi$ can be omitted without affecting the order of the error terms in the system (4.35). We can therefore replace (4.37) with a new fast quantity

$$\int e^{i(\omega y+\xi t)}\varphi^2\left\{\left[(\tfrac{\omega}{\xi})\frac{\alpha}{\sqrt{2}} - \frac{1}{i\xi}(\frac{1}{c\sqrt{2}})(\alpha(\gamma_{12}+\gamma_{32})+\alpha_y(\alpha-c))\right]\hat{w}^{(1)}\right.$$

(4.38)

$$\left. + \hat{w}^{(2)} - \frac{1}{i\xi}(\frac{1}{4c})(\alpha+c)(\gamma_{11}-\gamma_{13}+\gamma_{31}-\gamma_{33})\hat{w}^{(3)}\right\} d\omega d\xi .$$

We need to find a condition which suppresses (4.38) at the boundary $x = 0$. If the coefficients are independent of y and t, then the bracketed quantity in the integrand is the Fourier transform of (4.38), give or take a factor φ^2, and we can accomplish what we want by setting this quantity equal to zero at the boundary. If we do this, clear denominators, and then invert the Fourier transform, we obtain

$$\frac{\partial w^{(2)}}{\partial t} + \frac{\alpha}{\sqrt{2}}\frac{\partial w^{(1)}}{\partial y} - \frac{1}{c\sqrt{2}}[\alpha(\gamma_{12}+\gamma_{32}) + \alpha_y(\alpha-c)]w^{(1)}$$

(4.39)

$$- (\frac{\alpha+c}{4c})(\gamma_{11}-\gamma_{13}+\gamma_{31}-\gamma_{33})w^{(3)} = 0 \quad \text{for} \quad x = 0 .$$

If the coefficients depend on y or t this derivation is not

valid. However, we can show that (4.39) is still useful in this case. Suppose that this condition holds, and write it in the simpler form

$$\frac{\partial w^{(2)}}{\partial t} + \frac{\alpha}{\sqrt{2}} \frac{\partial w^{(1)}}{\partial y} + F_1 w^{(1)} + F_3 w^{(3)} = 0 . \tag{4.40}$$

If we apply the operator having the symbol $\frac{\varphi^2}{i\xi}$ to (4.40), the result is

$$\begin{aligned} 0 &= \frac{\varphi^2}{i\xi} \circ \left[i\xi w^{(2)} + \left(i\omega \frac{\alpha}{\sqrt{2}} + F_1 \right) w^{(1)} + F_3 w^{(3)} \right] \\ &= \varphi^2 w^{(2)} + \varphi^2 \left[\left(\frac{\omega}{\xi} \right) \frac{\alpha}{\sqrt{2}} + \frac{F_1}{i\xi} + \mathcal{O}\left(\frac{\omega}{\xi^2} \right) + \mathcal{O}\left(\frac{1}{\xi^2} \right) \right] w^{(1)} \\ &\quad + \varphi^2 \left[\frac{F_3}{i\xi} + \mathcal{O}\left(\frac{1}{\xi^2} \right) \right] w^{(3)} \\ &= \int e^{i(\omega y + \xi t)} \varphi^2 \left[\left\{ \left(\frac{\omega}{\xi} \right) \frac{\alpha}{\sqrt{2}} + \frac{F_1}{i\xi} \right\} \hat{w}^{(1)} + \hat{w}^{(2)} + \frac{F_3}{i\xi} \hat{w}^{(3)} \right] d\omega d\xi \\ &\quad + \mathcal{O}\left(\frac{\omega}{\xi} \varphi^2 \right) w + (\text{order}(-2)) w . \end{aligned} \tag{4.41}$$

According to (4.38) and the definitions of F_1 and F_3 implied by (4.40), the integral in the last line is our approximation to the incoming fast part of the solution. The entire last line is equal to zero, so this fast part must be equal to

$$\mathcal{O}\left(\frac{\omega}{\xi^2} \varphi^2 \right) w + \text{order}(-2) w$$

at $x = 0$. The incoming fast part is therefore small compared to w for large frequencies and for angles near normal incidence.

The boundary condition (4.39) is written in terms of the components of the vector w which appears in the system (4.3). We can use the definition (4.4) of w to write the condition in terms of the variables u, v, and p in the original system (4.1). When we do this the result is

$$\frac{\partial}{\partial t}(u+p) + \alpha\frac{\partial v}{\partial y} - \frac{1}{c}[\alpha(\gamma_{12}+\gamma_{32}) + \alpha_y(\alpha-c)]v$$

$$- \left(\frac{\alpha+c}{4c}\right)(\gamma_{11}-\gamma_{13}+\gamma_{31}-\gamma_{33})(u-p) = 0 \qquad (4.40)$$

$$\text{for } x = 0.$$

It may be worthwhile to compare (4.40) with the boundary condition

$$w^{(2)} = u + p = 0 \qquad (4.41)$$

which we mentioned early in Section 4.1. This condition was derived from the system (4.3), in which the coefficient of the x-derivative is a diagonal matrix. The newer condition (4.40) is based on the incoming fast variable (4.37) which was obtained from a more extensive uncoupling of the system. An inspection of (4.37) shows that this variable can be considered a perturbation of $w^{(2)} = u + p$, so in some sense (4.40) is a refinement of (4.41). One obvious difference between the two is the presence of derivatives in (4.40). This is a result of the need to clear denominators in the Fourier transform of the incoming fast part. The other difference is the presence of terms which do not involve $u + p$. The term $\alpha\frac{\partial v}{\partial y}$ is a result of uncoupling the leading symbol, and the other terms in (4.40) are the

result of reducing the coupling caused by terms of order zero. The term $\lambda_y(\alpha-c)$ corresponds to the part of the zero-order coupling which resulted from the prior uncoupling of the leading symbol. If we had not carried out the lower-order uncoupling, then the boundary condition would have been $\frac{\partial}{\partial t}(u+p) + \alpha \frac{\partial v}{\partial y} = 0$.

Up to now we have discussed boundary conditions only for the incoming fast part of the solution. If the boundary $x = 0$ is an inflow boundary, i.e., if $\alpha < 0$ in (4.3), then we must also specify a condition for the slow part. One possibility is to use the system (4.3) to obtain the condition

$$w^{(1)} = v = \text{given function}, \quad \text{for } x = 0. \tag{4.42}$$

Another possibility is to try to base a boundary condition on the more extensively uncoupled system (4.35). We could presumably prescribe a value for the Fourier transform of the slow component in (4.35), clear denominators, and then apply an inverse transform to obtain an inhomogeneous boundary condition analogous to (4.40).

The first approach suggested here is acceptable, but the second one is not. Our use of the cutoff function φ means that we have uncoupled the system only on the wedge Γ in the (ω,ξ) space which corresponds to rapidly moving waves. This is clearly no restriction when we are seeking boundary conditions which suppress the incoming fast part of the solution. But in the present case it is a major restriction, since the slow part of the solution is associated with the entire (ω,ξ) space. The partially uncoupled system (4.35)

cannot give a full description of the slow part, so there is no point in trying to use this system to find an improvement of the condition (4.42).

We will therefore prescribe the conditions

$$\text{(a)} \quad \frac{\partial}{\partial t}(u+p) + \alpha \frac{\partial v}{\partial y} - \frac{1}{c}[\alpha(\gamma_{12}+\gamma_{32}) + \alpha_y(\alpha-c)]v - \left(\frac{\alpha+c}{4c}\right)(\gamma_{11}-\gamma_{13}+\gamma_{31}-\gamma_{33})(u-p) = 0 \tag{4.43}$$

(b) v = given function, for $x = 0$,

if $x = 0$ defines an inflow boundary. In the case of an outflow boundary we will use only the first condition.

We need to discuss whether these conditions define a well-posed initial-boundary value problem. We will first consider the special case in which the zero-order terms in (4.43)(a) are not present. This would be the situation if we were to uncouple the leading symbol but do nothing about the zero-order coupling in the system. In this case the boundary conditions at inflow have the form

$$v = g$$
$$u+p = (u+p)_{t=0} - \int_0^t \alpha \frac{\partial g}{\partial y}(y,\tau)d\tau, \quad \text{for} \quad x = 0,$$

where g is a given function of y and t. Simple energy estimates for the system (4.3) show that this defines a well-posed problem. The a priori estimates for the solution involve a time integral of a tangential derivative of the data.

In the general case we must do something different in order to properly handle the incoming fast part of the solution. This is really no problem, since in principle it has already been done. In Section 2.6 we described a process for estimating the incoming fast part for systems in one space dimension, and at the end of Section 3.2 we indicated that this process extends to the multi-dimensional case with little modification. These estimates were derived in order to show that our boundary conditions are effective at suppressing the incoming fast part. One would expect that they also imply that the conditions give well-posed problems.

4.3 Effects of Orthogonal Changes of Spatial Coordinates

The spatial domain considered in the previous sections was the half-plane for which $x > 0$. In order to treat slightly more general regions we will now consider the effects of linear orthogonal changes of coordinates. We will first derive some general formulas, and we will then use these formulas to derive boundary conditions for the four sides of the unit square $0 < x < 1$, $0 < y < 1$. These conditions will be used for the test problem which will be discussed in the next section.

We first need to establish some notation. Let R denote an orthogonal transformation on $\mathbb{R}^2$, i.e., either a rotation or a flip of coordinates. This is illustrated in Figure 4.1 for the case where

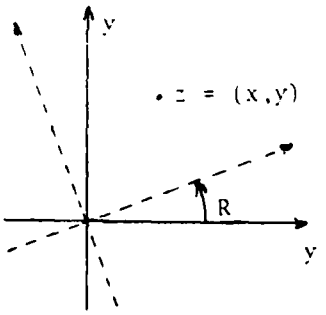


Figure 4.1

R is a rotation. Suppose that f is a function which is defined with respect to the old coordinates (solid axes), and let $\tilde{f}$ be a function defined in the new coordinates (dotted axes). From the figure we can see that if f is evaluated at point $z = (x,y)$, then the numbers plugged into $\tilde{f}$ must be given by $\tilde{z} = R^{-1}z$. We therefore have the relation

$$f(z) = \tilde{f}(\tilde{z}) = \tilde{f}(R^{-1}z) \tag{4.44}$$

for scalar functions. For vector-valued functions the dependent variables must also be transformed. In this case the vector $f = (f_1, f_2)^T$ describes a direction relative to the solid axes, and the vector $\tilde{f}$ describes the same direction relative to the dotted axes. The correct change of coordinates is therefore given by

$$f(z) = R\tilde{f}(R^{-1}z) . \tag{4.45}$$

We will need to use the fact that the gradient of a function and the divergence of a vector field are invariant in the sense implied by (4.44) and (4.45). We will give short proofs of these in order to help establish our notation.

First consider the gradient. If $f(z) = \tilde{f}(R^{-1}z)$, then

$$f'(z) = \tilde{f}'(R^{-1}z)R^{-1} ,$$

or

$$(\partial_1 f, \partial_2 f) = (\partial_1 \tilde{f}, \partial_2 \tilde{f})R^{-1} . \tag{4.46}$$

Here the subscripts 1 and 2 denote differentiation with respect to the first and second arguments, respectively. (4.46) can be written

$$\begin{pmatrix} \partial_1 f(z) \\ \partial_2 f(z) \end{pmatrix} = R \begin{pmatrix} \partial_1 \tilde{f}(R^{-1}z) \\ \partial_2 \tilde{f}(R^{-1}z) \end{pmatrix} . \tag{4.47}$$

Here we have used the fact that the transformation is orthogonal, i.e., $R^{-1} = R^T$. The invariance of the gradient follows from the observation

that (4.47) is a special case of the formula for changes of coordinates given in (4.45).

We now consider the divergence. If f and $\tilde{f}$ are vector-valued functions such that $f(z) = R\tilde{f}(R^{-1}z)$, then $f'(z) = R\tilde{f}'(R^{-1}z)R^{-1}$, or

$$(\partial_1 f, \partial_2 f) = R(\partial_1 \tilde{f}, \partial_2 \tilde{f})R^{-1}.$$

In this case $(\partial_1 f, \partial_2 f)$ is the Jacobian matrix of f. The divergence of f, $\partial_1 f_1 + \partial_2 f_2$, is equal to the trace of this Jacobian. The fact that the trace of a matrix is invariant under similarity transformations implies

$$(\operatorname{div} f)(z) = (\operatorname{div} \tilde{f})(R^{-1}z).$$

We will now study the effect of the transformation R on the system (4.1),

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ v \\ p \end{pmatrix} = \begin{pmatrix} \alpha & & -c \\ & \alpha & \\ -c & & \alpha \end{pmatrix} \frac{\partial}{\partial x}\begin{pmatrix} u \\ v \\ p \end{pmatrix} + \begin{pmatrix} \beta & & \\ & \beta & -c \\ & -c & \beta \end{pmatrix} \frac{\partial}{\partial y}\begin{pmatrix} u \\ v \\ p \end{pmatrix} + (\gamma_{ij})\begin{pmatrix} u \\ v \\ p \end{pmatrix}.$$

In order to make it fit our notation for changing coordinates, we will write this system in the form

$$W_t(z,t) = (W_1, W_2)\begin{pmatrix} \alpha \\ \beta \end{pmatrix} - c\begin{pmatrix} p_1 \\ p_2 \\ 0 \end{pmatrix} - c\begin{pmatrix} 0 \\ 0 \\ u_1 + v_2 \end{pmatrix} + (\gamma_{ij})W, \tag{4.48}$$

where $W = (u,v,p)^T$. The numerical subscripts on u, v, p, and W denote partial derivatives. The 3×2 matrix (W_1, W_2) is the

Jacobian matrix of W with respect to the spatial variables. It can be denoted by W'.

We will define the change of coordinates by

$$W(z,t) = \begin{pmatrix} R & \\ & 1 \end{pmatrix} \tilde{W}(R^{-1}z,t) . \tag{4.49}$$

The matrix in (4.49) is a 3×3 matrix in which the 2×2 matrix appears in the upper left. This matrix is present because it is necessary to transform the velocity components when the spatial coordinates are changed. If (4.49) is inserted into (4.48), the result is

$$\begin{aligned} \begin{pmatrix} R & \\ & 1 \end{pmatrix} \tilde{W}_t(R^{-1}z,t) = \begin{pmatrix} R & \\ & 1 \end{pmatrix} \tilde{W}'(R^{-1}z,t)\, R^{-1} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \\ - c\begin{pmatrix} p_1 \\ p_2 \\ 0 \end{pmatrix} - c\begin{pmatrix} 0 \\ 0 \\ u_1 + v_2 \end{pmatrix} + (\gamma_{ij}) \begin{pmatrix} R & \\ & 1 \end{pmatrix} \tilde{W} \end{aligned} \tag{4.50}$$

In the first term on the right we used the chain rule to evaluate the Jacobian matrix W'. (4.50) can also be written

$$\begin{aligned} \tilde{W}_t = (\tilde{W}_1, \tilde{W}_2)\, R^{-1} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} - c\begin{pmatrix} R^{-1} & \\ & 1 \end{pmatrix} \begin{pmatrix} p_1(z,t) \\ p_2(z,t) \end{pmatrix} \\ - c\begin{pmatrix} 0 \\ 0 \\ u_1 + v_2 \end{pmatrix} + \begin{pmatrix} R^{-1} & \\ & 1 \end{pmatrix} (\gamma_{ij}) \begin{pmatrix} R & \\ & 1 \end{pmatrix} \tilde{W} . \end{aligned} \tag{4.51}$$

According to the formula (4.47) for the invariance of the gradient, the second term on the right is $c(\tilde{p}_1(R^{-1}z,t), \tilde{p}_2, 0)^T$. The form of the

third term is invariant under the transformation, since $u_1 + v_2$ is the divergence of the velocity field. The system (4.51) is therefore the same as

$$\tilde{W}_t = (\tilde{W}_1, \tilde{W}_2)\begin{pmatrix}\tilde{\alpha}\\ \tilde{\beta}\end{pmatrix} - c\begin{pmatrix}\tilde{p}_1\\ \tilde{p}_2\\ 0\end{pmatrix} - c\begin{pmatrix}0\\ 0\\ \tilde{u}_1 + \tilde{v}_2\end{pmatrix} + (\tilde{\gamma}_{ij})\tilde{W}, \tag{4.52}$$

where

$$\begin{pmatrix}\tilde{\alpha}\\ \tilde{\beta}\end{pmatrix} = R^{-1}\begin{pmatrix}\alpha\\ \beta\end{pmatrix}$$

and (4.53)

$$(\tilde{\gamma}_{ij}) = \begin{pmatrix}R^{-1} & \\ & 1\end{pmatrix}(\gamma_{ij})\begin{pmatrix}R & \\ & 1\end{pmatrix}.$$

Equation (4.53) represents the transformation of the velocity field of the flow about which the system has been linearized.

Equation (4.52) shows that the form of the system (4.48) does not change under orthogonal transformations of spatial coordinates. This result depends on the fact that spatial derivatives appear only as gradients and divergences. The value of this result is that our previous calculations immediately give boundary conditions along any straight boundary. Suppose that our spatial domain is the region to the upper right of the line ℓ in Figure 4.2. Choose a coordinate system so that the $\tilde{y}$ axis coincides with ℓ and so that the positive $\tilde{x}$ direction points into the region. We can now apply our earlier calculations regarding boundary conditions which suppress the incoming fast part of the solution.

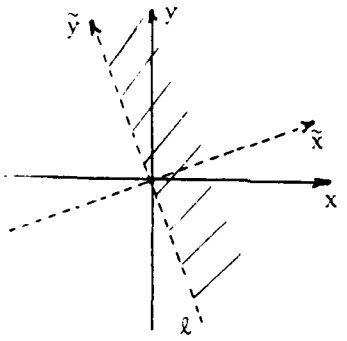


Figure 4.2

If this is an inflow boundary, then (4.43) implies that we can use

$$\frac{\partial}{\partial t}(\tilde{u}+\tilde{p}) + \tilde{\alpha}\frac{\partial \tilde{v}}{\partial \tilde{y}} - \frac{1}{c}\left[\tilde{\alpha}(\tilde{\gamma}_{12}+\tilde{\gamma}_{32}) + \tilde{\alpha}_{\tilde{y}}(\tilde{\alpha}-\tilde{c})\right]\tilde{v}$$

$$-\left(\frac{\tilde{\alpha}+c}{4c}\right)\left(\tilde{\gamma}_{11}-\tilde{\gamma}_{13}+\tilde{\gamma}_{31}-\tilde{\gamma}_{33}\right)(\tilde{u}-\tilde{p}) = 0 \qquad (4.54)$$

$\tilde{v}$ = given function, for $\tilde{x} = 0$.

If this is an outflow boundary, then we would use only the first condition. The conditions (4.54) can be expressed in terms of the coefficients and components of the original system (4.1) through the relations

$$z = \begin{pmatrix} x \\ y \end{pmatrix} = R\begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} = R\tilde{z}$$

$$\begin{pmatrix} u \\ v \end{pmatrix} = R\begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix}$$

$$p = \tilde{p}$$

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = R\begin{pmatrix} \tilde{\alpha} \\ \tilde{\beta} \end{pmatrix}$$

$$(\gamma_{ij}) = \begin{pmatrix} R & \\ & 1 \end{pmatrix}(\tilde{\gamma}_{ij})\begin{pmatrix} R^{-1} & \\ & 1 \end{pmatrix}.$$

We will now use these general formulas to derive boundary conditions for the sides of the unit square $0 < x < 1$, $0 < y < 1$. These conditions will be needed for the test problem which will be discussed in the next section.

Denote the sides of the squares in the manner indicated in Figure 4.3. For segment A we can use the conditions (4.43) which

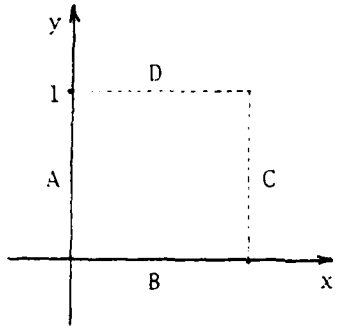


Figure 4.3

were derived earlier since this part of the boundary corresponds to $x = 0$. The inward normal direction for segment B is the positive y direction, so for this segment we need to use the transformation $\tilde{x} = y$, $\tilde{y} = x$. The matrix R is then given by

$$R_B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

For segment C the transformation is $\tilde{x} = -x$, $\tilde{y} = y$, and for segment D it is $\tilde{x} = -y$, $\tilde{y} = x$. The matrices for these transformations are

$$R_C = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad R_D = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

respectively.

Routine calculations produce the following boundary conditions:

Segment A (x = 0):

$$\frac{\partial}{\partial t}(u+p) + \alpha \frac{\partial v}{\partial y} - \frac{1}{c}[\alpha(\gamma_{12}+\gamma_{32}) + \alpha_y(\alpha-c)]v$$

(4.55)(A)
$$- \left(\frac{\alpha+c}{4c}\right)(\gamma_{11}-\gamma_{13}+\gamma_{31}-\gamma_{33})(u-p) = 0$$

v = given function.

Segment B (y = 0):

$$\frac{\partial}{\partial t}(v+p) + \beta \frac{\partial u}{\partial x} - \frac{1}{c}[\beta(\gamma_{21}+\gamma_{31}) + \beta_x(\beta-c)]u$$

(4.55)(B)
$$- \left(\frac{\beta+c}{4c}\right)(\gamma_{22}-\gamma_{23}+\gamma_{32}-\gamma_{33})(v-p) = 0$$

u = given function.

Segment C (x = 1):

$$\frac{\partial}{\partial t}(-u+p) - \alpha \frac{\partial v}{\partial y} - \frac{1}{c}[-\alpha(-\gamma_{12}+\gamma_{32}) - \alpha_y(-\alpha-c)]v$$

(4.55)(C)
$$- \left(\frac{-\alpha+c}{4c}\right)(\gamma_{11}+\gamma_{13}-\gamma_{31}-\gamma_{33})(-u-p) = 0$$

v = given function.

Segment D (y = 1):

$$\frac{\partial}{\partial t}(-v+p) - \beta\frac{\partial u}{\partial x} - \frac{1}{c}[-\beta(-\gamma_{21}+\gamma_{31}) - \beta_x(-\beta-c)]u$$

(4.55)(D)
$$- \left(\frac{-\beta+c}{4c}\right)(\gamma_{22}+\gamma_{23}-\gamma_{32}-\gamma_{33})(-v-p) = 0$$

u = given function.

For each segment the first condition is the one which is intended to suppress the incoming fast part of the solution. The second condition prescribes a value for the slow variable, and it should be imposed only when the segment is an inflow boundary.

4.4 Numerical Computations

In this section we present the results of some numerical computations involving the boundary conditions which have just been derived. We consider the system

$$(4.56)\quad \frac{\partial}{\partial t}\begin{pmatrix} u \\ v \\ p \end{pmatrix} = \begin{pmatrix} -1 & & -3 \\ & -1 & \\ -3 & & -1 \end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} u \\ v \\ p \end{pmatrix} + \begin{pmatrix} 0 & & \\ & 0 & -3 \\ & -3 & 0 \end{pmatrix}\frac{\partial}{\partial y}\begin{pmatrix} u \\ v \\ p \end{pmatrix} + (\gamma_{ij})\begin{pmatrix} u \\ v \\ p \end{pmatrix}$$

on the unit square $0 < x < 1$, $0 < y < 1$. This system is a special case of the system (4.1). We will use two different choices for the matrix (γ_{ij}).

We wish to compare three types of boundary conditions for this system. The first of these is obtained by diagonalizing the coefficient of the normal derivative and then defining boundary conditions in terms of the dependent variables in the new system. These variables will be referred to as "characteristic variables". This was discussed early in Section 4.1, immediately after equation (4.4). For the four sides of the unit square the incoming fast characteristic variables are the quantities in (4.55) which are differentiated with respect to time. The second set of conditions is obtained by uncoupling the leading symbol in the manner described earlier, but then doing nothing about the zero-order coupling in the system. These conditions can be obtained by deleting the zero-order terms in the derivative conditions appearing in (4.55). The third set of boundary conditions is obtained by also uncoupling the zero-order terms in the system of differential equations. These are the conditions (4.55).

We present two separate tests of these conditions, one to demonstrate the effect of uncoupling the leading symbol, and the other to demonstrate the effect of uncoupling both the leading symbol and the zero-order term. In the first case we let $\gamma_{ij} = 0$ for all i,j and use the first two sets of boundary conditions. In the second case we use all three sets of conditions, and we let (γ_{ij}) be the matrix

$$\begin{pmatrix} 0 & 10 & 0 \\ -10 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{4.57}$$

In the computations we set the solution equal to zero when $t = 0$. At the boundary $x = 0$ we set v (see (4.55)(A)) equal to a pulse consisting of half a sine wave in t multiplied by half a sine wave in the tangential variable y. We use homogeneous conditions on the other boundaries. The nonzero part of the solution is due entirely to the nonzero data at the boundary $x = 0$, so it is possible to study the influence of these data by examining the size of the solution in various parts of the (x,y) plane at various times.

In our computation the system is approximated by the leap frog difference scheme. The derivative boundary conditions in (4.55) are approximated by centered differences in the time and tangent variables. The outgoing fast characteristic variables are extrapolated at the boundary using the given differential equation. For this we use centered differences in the time and tangent variables, and we approximate the normal derivative with a forward difference which uses a time

average at the back point. At an outflow boundary the slow characteristic variable is extrapolated in the same manner.

The boundaries $y = 0$ and $y = 1$ are characteristic for the system (4.56). At these boundaries we integrate the slow characteristic variable in the boundary using a centered difference approximation. This is an experiment to see if the incoming fast modes can be activated at a characteristic boundary. In our earlier discussion we always assumed that our boundary was noncharacteristic.

The surfaces pictured in Figures 4.5 and 4.6 are graphs of $(u^2 + v^2 + p^2)^{1/2}$ as functions of x and y for fixed t. The configuration is shown in Figure 4.4.

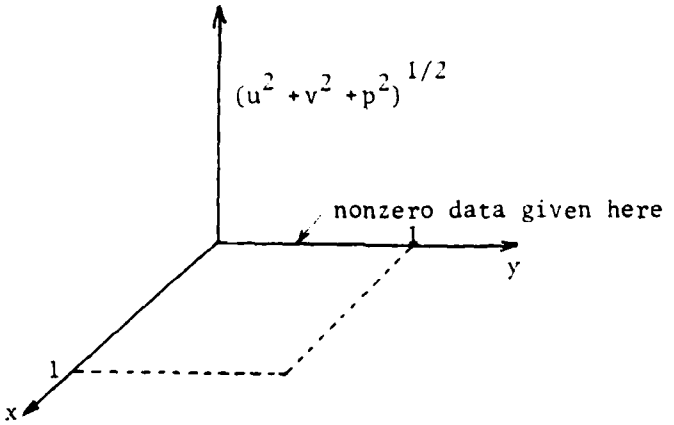


Figure 4.4

We show solutions at times $t = .125$, $.25$, and $.375$. The fast mode entering through the boundary $x = 0$ has normal velocity 4 since $\alpha = -1$ and $c = -3$. Pulses entering on this mode should therefore

be visible near the nearest boundary ($x = 1$) in the graphs for $t = .25$.

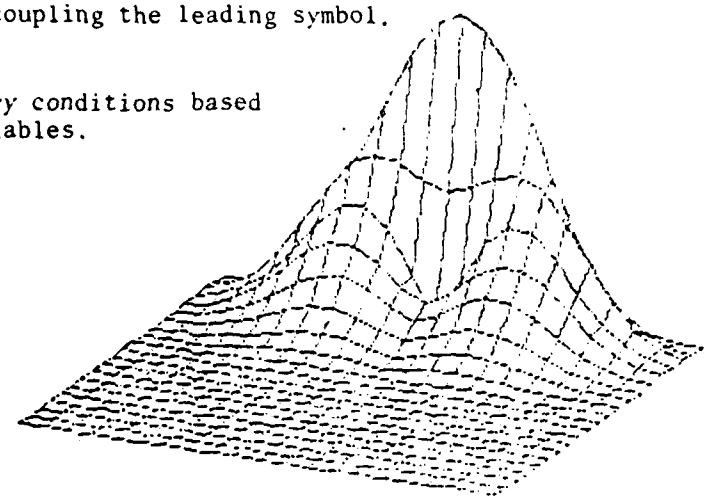
In Figure 4.5 we show the effect of uncoupling the leading symbol. In this case $\gamma_{ij} = 0$ for all i,j. Figure 4.5(a) shows the solution corresponding to the boundary conditions defined in terms of characteristic variables. The solution in Figure (4.5)(b) corresponds to the more refined boundary conditions. The second set of conditions is clearly more effective at suppressing the incoming fast part of the solution.

In Figure 4.6 we show the effect of uncoupling both the leading symbol and the term of order zero. In this case the matrix (γ_{ij}) is given by (4.57). The simplest boundary conditions are used in part (a). In part (b) we use the boundary conditions obtained by uncoupling the leading symbol only. The boundary conditions for part (c) are obtained by uncoupling both the leading symbol and the term of order zero. The third set of conditions is clearly the most effective.

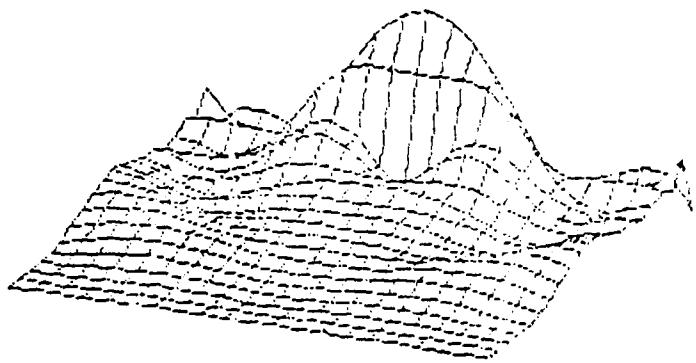
Figure 4.5. Effect of uncoupling the leading symbol.

(a) Solution using boundary conditions based on characteristic variables.

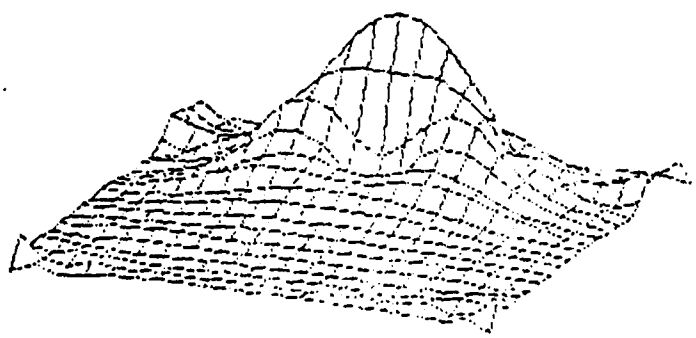
t = .125



t = .25

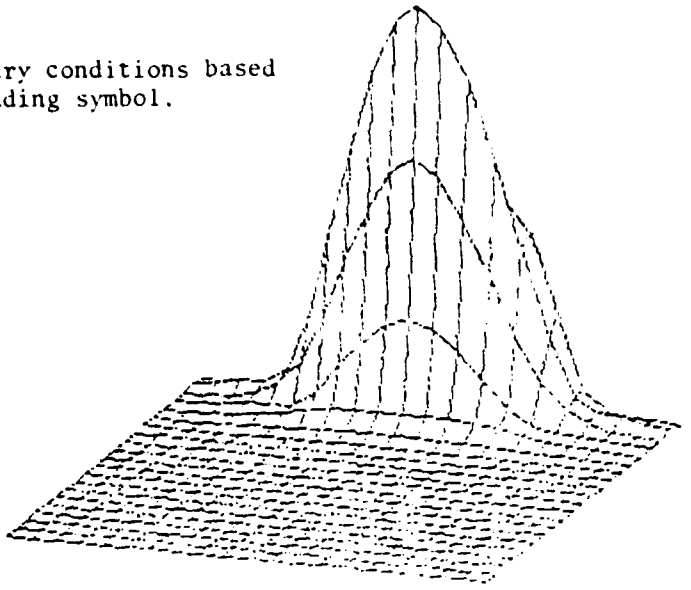


t = .375

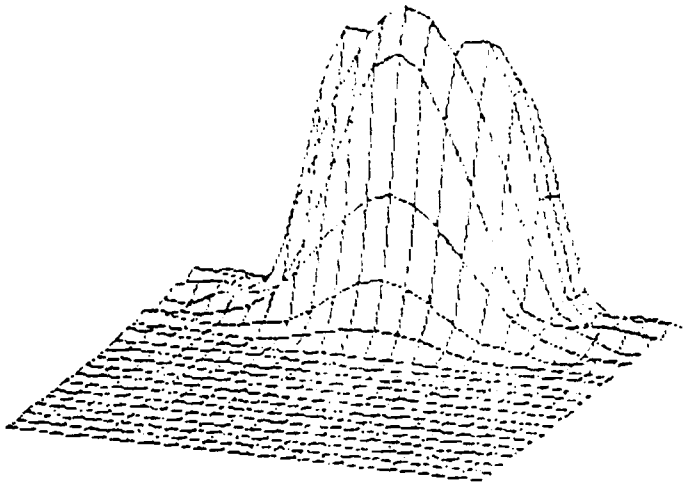


(b) Solution using boundary conditions based on uncoupling the leading symbol.

t = .125



t = .25



t = .375

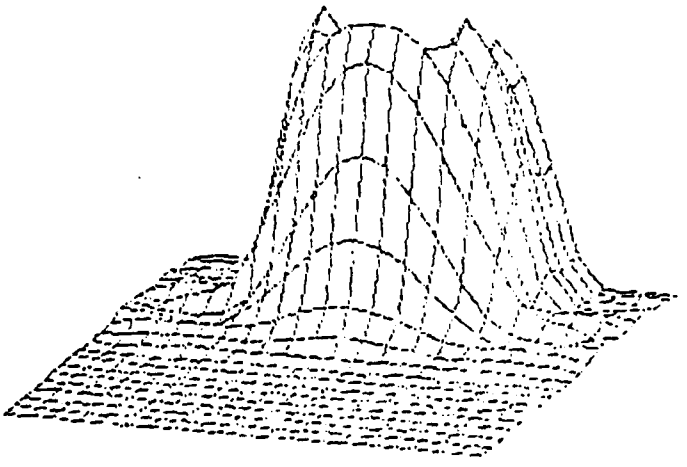
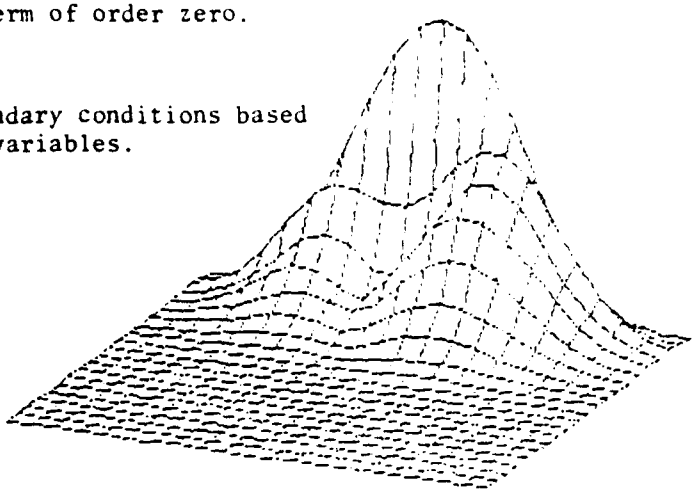


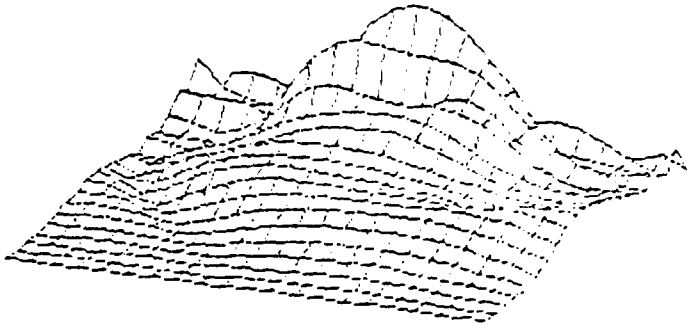
Figure 4.6. Effect of uncoupling both the leading symbol and the term of order zero.

(a) Solution using boundary conditions based on characteristic variables.

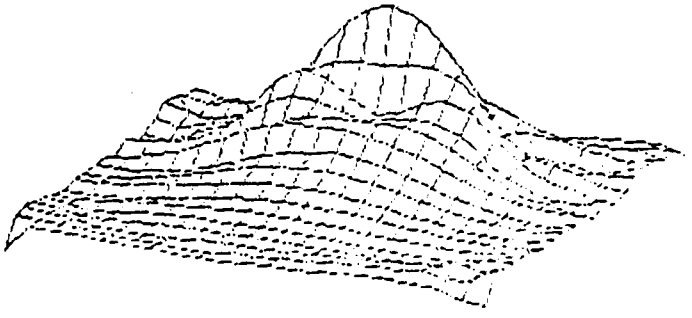
t = .125



t = .25

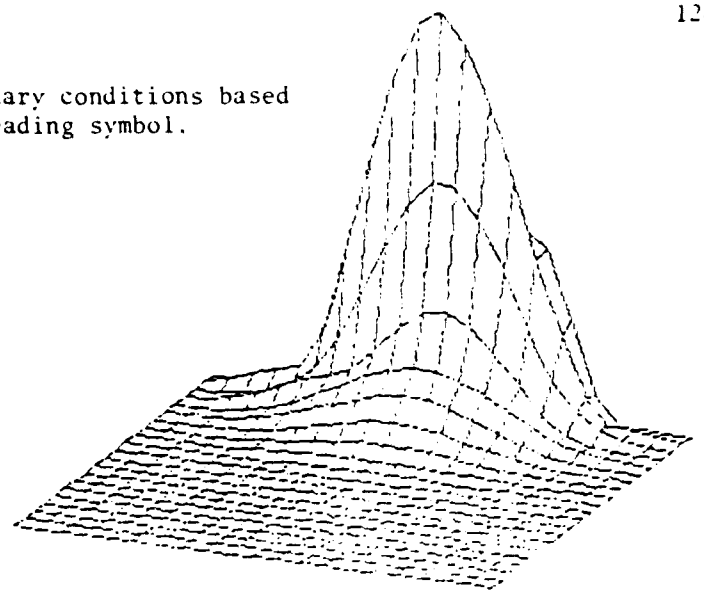


t = .375

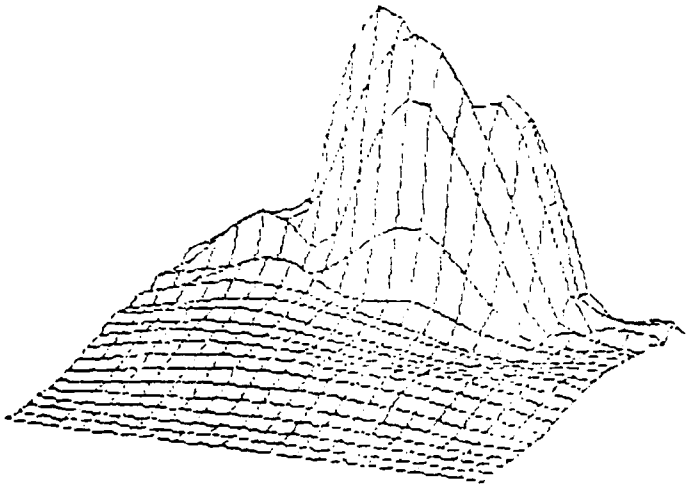


(b) Solution using boundary conditions based on uncoupling the leading symbol.

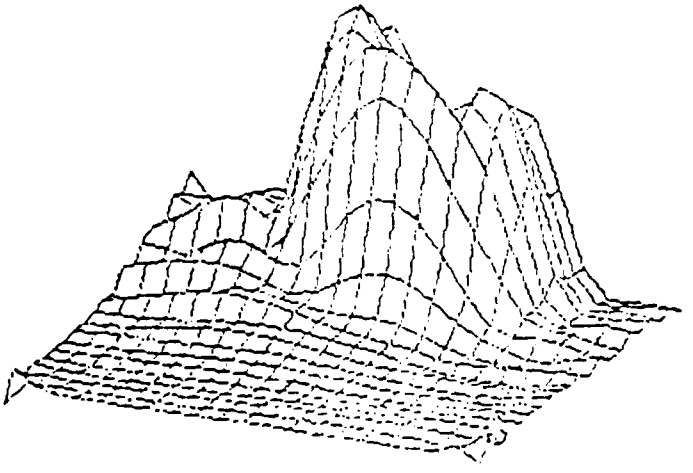
t = .125



t = .25

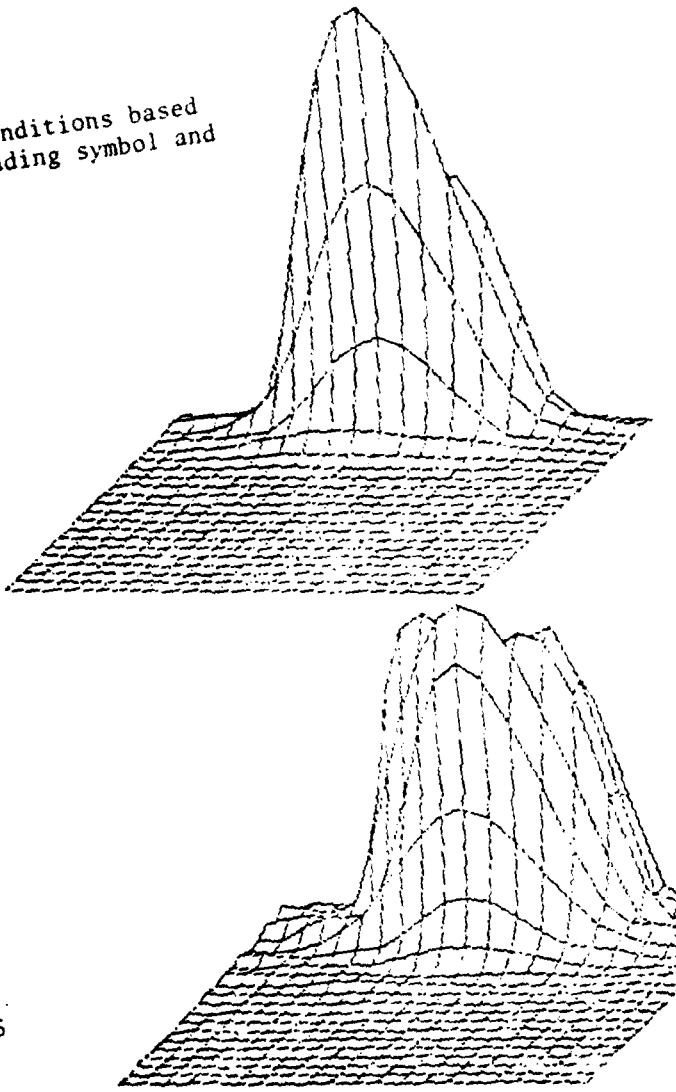


t = .375

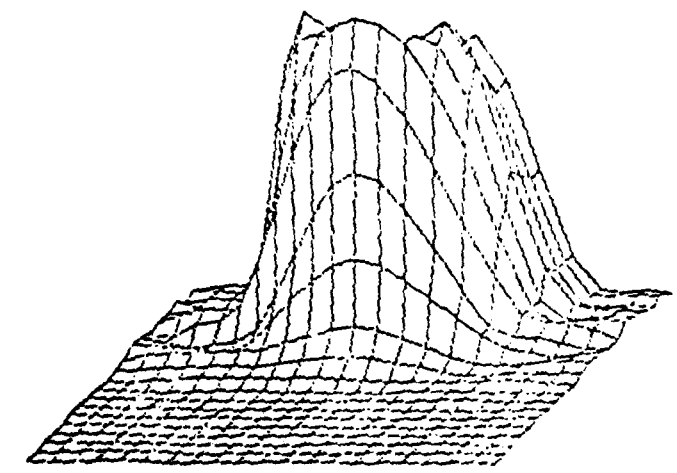


(c) Solution using boundary conditions based on uncoupling both the leading symbol and the term of order zero.

t = .125



t = .25



t = .375

APPENDIX

PROPERTIES OF PSEUDO DIFFERENTIAL OPERATORS

In this appendix we will define pseudo differential operators and state without proof of some of their basic properties. More extensive treatments can be found in Nirenberg [6], Taylor [9], and Treves [1 1].

We must first establish some notation. Partial derivatives in $\mathbb{R}^n$ will be denoted by ∂^α, where $\alpha = (\alpha_1,\ldots,\alpha_n)$, and

$$\partial^\alpha = \partial_1^{\alpha_1} \cdot \ldots \cdot \partial_n^{\alpha_n} = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdot \ldots \cdot \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n} .$$

The α_j are nonnegative integers. Differential operators can then be written in the form

$$P = \Sigma\, a_\alpha \partial^\alpha .$$

The coefficients a_α are functions on $\mathbb{R}^n$, and the sum is taken over finitely many multi-indices α. We will allow the possibility that P may act on vector-valued functions. In that case the a_α may be either scalars or matrices.

The Fourier transform on $\mathbb{R}^n$ will be denoted by

$$\hat{u}(\xi) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-ix\cdot\xi}\, u(x)dx ,$$

where $x \cdot \xi = \Sigma_{j=1}^n x_j\xi_j$. The inverse Fourier transform is then given by

$$\int_{\mathbb{R}^n} e^{ix\cdot\xi}\, \hat{u}(\xi)d\xi .$$

Differential operators can be represented in terms of the Fourier transform. For suitable functions u, we have

$$\begin{aligned}(Pu)(x) &= \Sigma\, a_\alpha(x)\partial^\alpha u(x) \\ &= \Sigma\, a_\alpha(x)\partial_x^\alpha \int e^{ix\cdot\xi}\, \hat{u}(\xi)d\xi \\ &= \int e^{ix\cdot\xi}\, [\Sigma\, a_\alpha(x)(i\xi)^\alpha]\, \hat{u}(\xi)d\xi .\end{aligned}$$

Here $(i\xi)^\alpha$ denotes the product $(i\xi_1)^{\alpha_1}\cdot\ldots\cdot(i\xi_n)^{\alpha_n}$. The equation can be written in the form

$$(Pu)(x) = \int e^{ix\cdot\xi}\, p(x,\xi)\hat{u}(\xi)d\xi , \tag{A.1}$$

where $p(x,\xi) = \Sigma\, a_\alpha(x)(i\xi)^\alpha$. The function p is sometimes called the symbol of the operator P.

Pseudo differential operators are obtained by allowing a larger class of symbols to be used in (A.1). Every differential operator is a pseudo differential operator, but not vice versa. One fairly general symbol class is the class $S^m_{\rho,\delta}$, $0 \le \delta < \rho \le 1$, which was introduced by Hörmander. This is defined to be the set of all C^∞ functions p which satisfy estimates of the form

$$|\partial_\xi^\alpha\, \partial_x^\beta\, p(x,\xi)| \le C_{K,\alpha,\beta}(1+|\xi|)^{m-\rho|\alpha|+\delta|\beta|} \quad ; \quad x \in K, \quad \xi \in \mathbb{R}^n \tag{A.2}$$

for all α,β and for every compact subset K of $\mathbb{R}^n$. The constant is allowed to depend on α, β, and K. The symbol of a differential operator of order m having smooth coefficients clearly belongs to the class $S^m_{1,0}$. This class will also be denoted by S^m. For the operators considered in this paper we always have $\rho = 1$ and $\delta = 0$. In general, the number m appearing in (A.2) is called the order of the operator P whose symbol is p. The order need not be positive, and it need not be an integer.

If $u \in C_0^\infty$, then $Pu \in C^\infty$. It is possible to extend P so that Pu is defined for any distribution u having compact support. In this case Pu is a distribution.

A useful concept is that of an asymptotic expansion of a symbol. Suppose that $\{m_j\}_{j=0}^\infty$ is a sequence of real numbers such that $m_j > m_{j+1}$ for all j and $m_j \to -\infty$ as $j \to \infty$. Let $\{p_j\}$ be a sequence of symbols such that $p_j \in S^{m_j}$ for each j. A symbol is said to be an asymptotic sum of the p_j, written

$$p \sim \sum_{j=0}^{\infty} p_j ,$$

provided $p - \Sigma_{j=0}^{k} p_j \in S^{m_{k+1}}$ for all k. That is, the error in each partial sum must have the same order as the first term omitted from the partial sum. This concept is analogous to the usual concept of asymptotic expansion. In fact, if a function $p(\xi)$ of one variable has an asymptotic expansion

$$p(\xi) \sim \sum_{j=0}^{\infty} \frac{a_j}{\xi^j} \quad \text{as} \quad \xi \to \infty$$

in the usual sense, then this expansion is also asymptotic in the sense described above.

Pseudo differential operators can be composed. Let P and Q be operators with symbols $p(x,\xi)$ and $q(x,\xi)$, respectively, and suppose that q has compact support in x. The composition $P(Qu)$ is then well-defined and is given by a pseudo differential operator whose symbol has the asymptotic expansion

$$\sigma_{PQ} \sim \sum_{|\alpha| \geq 0} \frac{1}{i^{|\alpha|}\alpha!} \partial_\xi^\alpha p(x,\xi) \partial_x^\alpha q(x,\xi) . \tag{A.3}$$

The sum is taken over all multi-indices $\alpha = (\alpha_1,\ldots,\alpha_n)$ having non-negative components. The order of α is given by $|\alpha| = \Sigma\alpha_j$, and the factorial $\alpha!$ denotes the product $\alpha_1!\alpha_2!\cdot\ldots\cdot\alpha_n!$.

It follows from (A.2) that when a symbol is differentiated with respect to ξ, the result is a symbol of lower order. This implies that the leading order term in (A.3) corresponds to $|\alpha| = 0$ and is equal to $p(x,\xi)q(x,\xi)$. The symbol of the product of two operators is therefore equal to the product of their symbols, up to certain terms of lower order.

This makes sense when we consider the special case of differential operators. The composition of two operators $a(x)\partial^\alpha$ and $b(x)\partial^\beta$ is equal to

$$a\partial^{\alpha}(b\partial^{\beta}) = ab\partial^{\alpha+\beta} + \text{lower order terms involving derivatives of } b.$$

In this case the composition law can be derived using Leibniz' rule. For general pseudo differential operators the derivation is much more complicated.

It is sometimes necessary to discuss adjoints of pseudo differential operators. The adjoint of an operator P is a pseudo differential operator P^* whose symbol has the asymptotic expansion

$$\sigma_{P^*} \sim \sum_{|\alpha| \geq 0} \frac{1}{i^{|\alpha|}\alpha!} \partial_{\xi}^{\alpha}\partial_{x}^{\alpha} p^*(x,\xi) .$$

The leading order term in this expansion is the adjoint of the symbol of P.

Bibliography

1. B. Engquist and A. Majda, "Absorbing boundary conditions for the numerical simulation of waves," Math. Comp. 31, 1977, pp. 629-651.

2. B. Engquist and A. Majda, "Radiation boundary conditions for acoustic and elastic wave calculations," Comm. Pure Appl. Math. 32, 1979, pp. 313-357.

3. L. Hörmander, "Linear differential operators," Proc. Internat. Congress Math. (Nice, 1970), Vol. 1, Gauthier-Villars, Paris, 1971, pp. 121-133.

4. H.-O. Kreiss, "Problems with different time scales for ordinary differential equations," SIAM J. Numer. Anal. 16, 1979, pp. 980-998.

5. H.-O. Kreiss, "Problems with different time scales for partial differential equations," Comm. Pure Appl. Math. 33, 1980, pp. 399-439.

6. L. Nirenberg, Lectures on Linear Partial Differential Equations, C.B.M.S. Regional Conf. Ser. in Math., No. 17, Amer. Math. Soc., Providence, R. I., 1973.

7. J. Oliger and A. Sundström, "Theoretical and practical aspects of some initial boundary value problems in fluid dynamics," SIAM J. Appl. Math. 35, 1978, pp. 419-446.

8. R. E. O'Malley, Jr. and L. R. Anderson, "Singular perturbations, order reduction, and decoupling of large scale systems," Numerical Analysis of Singular Perturbation Problems (Hemker and Miller, eds.), Academic Press, London, 1979, pp. 317-338.

9. M. Taylor, Pseudo Differential Operators, Lecture Notes in Mathematics, No. 416, Springer-Verlag, Berlin, 1974.

10. M. Taylor, "Reflections of singularities of solutions to systems of differential equations," Comm. Pure. Appl. Math. 28, 1975, pp. 457-478.

11. F. Treves, Introduction to Pseudodifferential Operators and Fourier Integral Operators (2 volumes), Plenum Press, New York, 1980.


ATE
LME